(54) Title: A SYSTEM AND A METHOD FOR OBJECT TRACKING

(57) Abstract: Disclosed is a method for object tracking, comprising: determining a region of interest (ROI) in a first frame of a video sequence, wherein the ROI is centered at a ground truth target location for objects to be tracked; feeding the determined ROI forward through a first CNN (convolutional network) to obtain a plurality of first feature maps in a higher layer of the CNN and a plurality of second feature maps in a lower layer of the first CNN, wherein the first CNN is pre-trained on an image classification task such that the first feature maps include more semantic features to determine a category for objects to be tracked in the video sequences, while the second feature maps carry more discriminative information to separate the objects from distractors with similar appearance; selecting a plurality of feature maps from the first and second feature maps, respectively by training a second CNN and a third CNN (sel-CNNs) with the first feature maps and the second feature maps, respectively; predicting, based on the selected first and second feature maps, two target heat maps indicating a target location for said objects in the current frame, respectively; and estimating, based on the two predicated target heat maps, a final target location for the object in the current frame.

Fig. 2
Technical Field

[0001] The disclosures relate to a system for object tracking and a method thereof.

Background

[0002] Visual tracking, as one of the fundamental problems in computer vision, has found wide applications. Although much progress has been made in the past decade, tremendous challenges still exist in designing a robust tracker that can well handle significant appearance changes, pose variations, severe occlusions, and background clutters.

[0003] In order to address these issues, existing appearance based tracking methods adopt either generative or discriminative models to separate the foreground from background and distinct co-occurring objects. One major drawback of these methods is that they rely on low-level hand-crafted features which are incapable to capture semantic information of targets, not robust to significant appearance changes, and only have limited discriminative power.

[0004] Driven by the emergence of large-scale visual data sets and fast development of computation power, Deep Neural Networks (DNNs), especially convolutional neural networks (CNNs), with their strong capabilities of learning feature representations, have demonstrated record breaking performance in image classification and object detection. Different from hand-crafted features, features learned by CNNs from massive annotated visual data and a large number of object classes (such as Image Net) carry rich high-level semantic information and are strong at distinguishing objects of different categories. These features have good generalization capability across data sets. Recent studies have also shown that such features are robust to data corruption. Their neuron responses have strong selectiveness on object identities, i.e. for a particular object only a subset of neurons are
responded and different objects have different responding neurons.

Summary of Invention

[0005] The proposed solutions in the present application are based on the following observations.

[0006] First, CNN features at different levels/depths have different properties that fit the tracking problem. A top convolutional layer captures more abstract and high-level semantic features. They are strong at distinguishing objects of different classes and are very robust to deformation and occlusion. However, they are less discriminative to objects of the same category as shown by the examples. A lower layer provides more detailed local features which help to separate the target from distracters (e.g. other objects in the same class) with similar appearance. But they are less robust to dramatic change of appearance. Based on these observations, the present application is proposed to automatically switch the usage of these two layers during tracking depending on the occurrence of distracters.

[0007] Second, the CNN features pre-trained on Image Net are for distinguishing generic objects. However, for a particular target, not all the features are useful for robust tracking. Some feature responses may serve as noise. It is hard to distinguish the target object from background if all the feature maps are used. In contrast, through proper feature selection, the noisy feature maps not related to the representation of the target are cleared out and the remaining ones can more accurately highlight the target and suppress responses from background. The present application proposes a principled method to select discriminative feature maps and discard noisy or unrelated ones for the tracking target.

[0008] In one aspect of the present application, disclosed is a method for object tracking, comprising:

determining a region of interest (ROI) in a first frame of a video sequences, wherein the
ROI is centered at a ground truth target location for objects to be tracked;

feeding the determined ROI forward through a first CNN (convolutional network) to obtain a plurality of first feature maps in a higher layer of the CNN and a plurality of second feature maps in a lower layer of the first CNN, wherein the first CNN is pre-trained on image classification task such that the first feature maps include more semantic features to determine a category for objects to be tracked in the video sequences, while the second feature maps carry more discriminative information to separate the objects from distracters with similar appearance;

selecting a plurality of feature maps from the first and second feature maps, respectively by training a second CNN and a third CNN (sel-CNNs) with the first feature maps and the second feature maps, respectively;

predicting, based on the selected first and second feature maps, two target heat maps indicating a target location for said objects in the current frame, respectively; and

estimating, based on the two predicated target heat maps, a final target location for the object in the current frame.

[0009] In one aspect of the present application, disclosed is a system for object tracking, comprising:

a feature map selection unit configured to determine a region of interest (ROI) in a first frame of a video sequences, wherein the ROI is centered at a ground truth target location for objects to be tracked, and feed the determined ROI forward through a first CNN (convolutional network), and obtain a plurality of first feature maps in a higher layer of the CNN and a plurality of second feature maps in a lower layer of the first CNN;

a heat map prediction unit electrically communicated with the feature map selection unit and configured to predict, based on the first and the second feature maps, two target heat maps indicating a target location in the current frame, respectively; and

a target localization unit configured to, based on the two predicated heat maps, estimate a final target location for the ROI in the current frame,
wherein the first CNN is pre-trained on image classification task such that the first feature maps include more semantic features to determine a category for objects to be tracked in the video sequences, while the second feature maps carry more discriminative information to separate the objects from distracters with similar appearance.

[0010] In a further aspect of the present application, disclosed is a system for object tracking, comprising:

- a memory that stores executable components; and
- a processor that executes the executable components to perform operations of the system, the executable components may comprise:

  - a feature map selection component configured to determine a region of interest (ROI) in a first frame of a video sequences, wherein the ROI is centered at a ground truth target location for objects to be tracked, and feed the determined ROI forward through a first CNN (convolutional network), and obtain a plurality of first feature maps in a higher layer of the CNN and a plurality of second feature maps in a lower layer of the first CNN;

  - a heat map prediction component configured to predict, based on the first and the second feature maps, two target heat maps indicating a target location in the current frame, respectively; and

  - a target localization component configured to, based on the two predicated heat maps, estimate a final target location for the ROI in the current frame.

[0011] Compared to prior visual tracking algorithms, the proposed or claimed solutions have at least one of the two advantages which significantly benefit tracking performance:

1) A principled feature selection method is utilized to automatically select discriminative feature maps of convolutional networks and discard noisy or unrelated ones for more accurate visual tracking.

2) Two convolutional layers of different levels are jointly considered so that they complement with each other. Specifically, a top layer encodes more semantic features
and serves as a category detector, while a lower layer carries more discriminative information and can better separate the target from distracters with similar appearance.

**Brief Description of the Drawing**

[0012] Exemplary non-limiting embodiments of the present invention are described below with reference to the attached drawings. The drawings are illustrative and generally not to an exact scale. The same or similar elements on different figures are referenced with the same reference numbers.

[0013] Fig. 1 is a schematic scenario illustrating the operation of the system shown in Fig. 1 according to one embodiment of the present application.

[0014] Fig. 2 is a schematic diagram illustrating a system for object tracking applicable to schematic scenario of Fig. 1 according to an embodiment of the present application.

[0015] Fig. 3 is a schematic diagram illustrating a flow chart for a method for object tracking according to one embodiment of the present application.

[0016] Fig. 4 is a schematic diagram illustrating a flow chart for the training step S32 according to one embodiment of the present application.

[0017] Fig. 5 is a schematic diagram illustrating a flow chart for the predicting step S33 according to one embodiment of the present application.

[0018] Fig. 6 is a schematic diagram illustrating a flow chart for the predicting step S34 according to one embodiment of the present application.

[0019] Fig. 7 illustrates a system for object tracking consistent with an embodiment of the present application.

**Detailed Description**

[0020] Reference will now be made in detail to some specific embodiments of the invention including the best modes contemplated by the inventors for carrying out the invention. Examples of these specific embodiments are illustrated in the accompanying drawings. While the invention is described in conjunction with these specific embodiments,
it will be understood that it is not intended to limit the invention to the described embodiments. On the contrary, it is intended to cover alternatives, modifications, and equivalents as may be included within the spirit and scope of the invention as defined by the appended claims. In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present invention. The present invention may be practiced without some or all of these specific details. In other instances, well-known process operations have not been described in detail in order not to unnecessarily obscure the present invention.

[0021] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of the invention. As used herein, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms "comprises" and/or "comprising," when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

[0022] Fig. 1 is a schematic scenario illustrating the operation of the system for object tracking according to an embodiment of the present application. For a given target, a feature map selection method is performed on the a lower layers and a higher layer (for example, conv4-3 and conv5-3 layers, as shown) of the VGG network to select the most relevant feature maps and avoid over fitting on noisy ones. A general neural network (GNet) that captures the category information of the target is built on top of the selected feature maps of the higher layer (conv5-3 layer) of a fully convolutional network. A specific neural network (SNet) that discriminates the target from background with similar appearance is built on top of the selected feature maps of the lower (conv4-3) layer of the fully convolutional network. Both GNet and SNet are initialized in the first frame to perform foreground heat map
regression for the target and adopt different online update strategies. For a new frame, a region of interest (ROI) centered at the last target location containing both target and background context is cropped and propagated through the fully convolutional network such that two foreground heat maps are generated, by the GNet and the SNet, from the selected first feature maps from the higher layer of the fully convolutional network and the selected second feature maps from the lower layer of the fully convolutional network, respectively.

[0023] The SNet and the GNet share the same network architecture that consists of two additional convolutional layers. The first additional convolutional layer has convolutional kernels of size (for example, $9 \times 9$) and outputs a plurality of (for example, 36) feature maps as the input to the next layer. The second additional convolutional layer has kernels with a relative small size (for example, $5 \times 5$) and outputs the foreground heat map of the input image. ReLU is chosen as the nonlinear function for these two layers. Target localization is performed independently based on the two heat maps. The final target is determined by a distracter detection scheme that decides which heat map to be used.

[0024] Fig. 2 is a schematic diagram illustrating a system for object tracking applicable to schematic scenario of Fig. 1 according to an embodiment of the present application. As shown, the system 1000 for object tracking may comprise a feature map selection unit 10, a heat map prediction unit 20, and a target localization unit 30. Hereinafter, the feature map selection unit 10, a heat map prediction unit 20, and a target localization unit 30 will be further discussed, respectively.

**The feature map selection unit 10**

[0025] The feature map selection unit 10 operates to determine a region of interest (ROI) in a first frame of a video sequences, the ROI centered at a ground truth target location for the objects to be tracked, wherein the predetermined ground truth target location includes both the target region and background context for objects to be tracked, and feed each
determined ROI forward through a first CNN (convolutional network), and obtains a plurality of first feature maps in a higher layer of the CNN and a plurality of second feature maps in a lower layer of the CNN; wherein the CNN is pre-trained on image classification task such the first feature maps include more semantic features to determine a category for objects to be tracked in the video sequences, while the second feature maps carry more discriminative information to separate the objects from distracters with similar appearance.

[0026] The feature map selection unit 10 is configured to initialize the two convolutional neural networks with the first feature maps and the second feature maps, respectively, by inputting the two features maps into the two CNNs respectively to output a predicted location for the object in each CNN, comparing the predicted location with a ground truth target location for the object to obtain a prediction error; back-propagating the error through the CNN until the obtained error is less than a threshold, and output a plurality of feature maps from each of the CNNs. Then the feature map selection unit 10 operates to determine the significance of each feature map on an objective function, rank all the feature maps in a descending order according to their significance values, select the top ranked K feature maps from both higher and lower layers, wherein at the online tracking stage for the following frame, the feature maps are extracted and the corresponding K features maps are selected from the extracted maps, as will be discussed as below.

[0027] As can be seen, the proposed feature map selection is based on a target heat map regression model through the first CNN, also named as sel-CNN, and is conducted independently on, for example, the conv4-3 and conv5-3 layers of VGG. The sel-CNN (the first CNN) consists of a dropout layer followed by a convolutional layer without any nonlinear transformation. It takes the feature maps (conv4-3 or conv5-3) to be selected as input to predict the target heat map M, which is a 2-dimensional Gaussian centered at the ground truth target location with variance proportional to the target size. The sel-CNN is trained by minimizing the square loss between the predicted foreground heat map \( \hat{M} \) and the
After parameter learning using back-propagation converges, the selected CNN parameters are fixed and the feature maps are selected according to their impacts on the loss function. The input feature maps $F$ are vectorized into a vector denoted by $\text{vec}(F)$. Denote $f_i$ as the $i$-th element of $\text{vec}(F)$. The change of the loss function caused by the perturbation of the feature map $\delta F$ can be computed by a two-order Taylor expansion as follows:

$$
\delta L_{sel} = \sum_i g_i \delta f_i + \frac{1}{2} \sum_i h_{ii} (\delta f_i)^2 + \frac{1}{2} \sum_{i\neq j} h_{ij} \delta f_i \delta f_j
$$

(2)

where $g_i = \frac{dL_{sel}}{\delta f_i}$ and $h_{ij} = \frac{d^2 L_{sel}}{\delta f_i \delta f_j}$ are, respectively, the first and second order derivatives of the objective function with respect to the input feature maps. The number of elements in the feature maps is very large, for example, it will be more than 270,000 in some instances. The complexity for computing all the second order derivatives $h_{ij}$ is $O(270,000^2)$, which is too time consuming. Accordingly, in the present embodiment, the Hessian matrix with a diagonal matrix is approximated, in which the third term of the right hand side of formula (2) is neglected. Both the first derivatives $g_i$ and the second derivatives $h_{ij}$ can be efficiently computed via back-propagation.

The significance of the element $f_i$ is defined as the change of the objective function after setting $f_i$ to zero, i.e., $\delta f_i = 0 - f_i$. According to formula (2), the significance of $f_i$ can then be computed as

$$
\delta f_i = g_i f_i + \frac{1}{2} \delta f_i^2
$$

(3)
The significance of the $k$-th feature map is further defined as the summation of significance of all its elements

$$S_k = \sum_{x,y} s(x,y,k)$$

(4)

where $s(x,y,k)$ is the significance of the element indexed by location $(x,y)$ on the $k$-th feature map. All the feature maps are sorted in the descending order by their significance, and the top $K$ feature maps are selected for the current tracking task. These selected feature maps have significant impact on the objective function and thus are most relevant to the tracking task. Our feature map selection method can be conducted in an online fashion. In our experiments, we only conduct feature selection at the first frame and have achieved good performance. This should be partially attributed to the robustness of CNN features.

[0030] The idea of using quadratic approximation of the cost function to remove some connections in neural network can be traced back to year of 1989. The aim was to reduce the number of parameters and improve speed, while the present application targets on removing noisy feature maps and improving tracking accuracy.

**The heat map prediction unit 20**

[0031] The heat map prediction unit 20 receives the first and the second feature maps to predict, based on the received features maps, two target heat maps indicating the target location in the current frame, respectively. According to one embodiment of the present application, the heat map prediction unit 20 is configured with a general network and a specific network, wherein the general network (GNet) captures the category information of the target object based on top of the selected feature maps from the first feature maps; and the specific network (SNet) discriminates the target object from background with similar appearance based on top of the selected second feature maps of the conv4-3 layer. As discussed in the above, the SNet and The GNet share the same network architecture that consists of two additional convolutional layers. The first additional convolutional layer has convolutional kernels of size (for example, $9 \times 9$) and outputs a plurality of (for example, 36)
feature maps as the input to the next layer. The second additional convolutional layer has
kernels with a relative small size (for example, $5 \times 5$) and outputs the foreground heat map of
the input image. ReLU is chosen as the nonlinear function for these two layers.

[0032] Both GNet and SNet are initialized in the first frame to perform foreground heat
map regression for the target object. The SNet and the GNet are initialized in the first frame
by minimizing the following square loss function:

$$ L = L_s + L_g $$

$$ L_U = \| \hat{M}_U - M \|^2_F + \beta \| W_U \|^2_F $$

(5)

where the subscript $ U \equiv \{ S, G \} $ indicates SNet and GNet, respectively; $ \hat{M}_U $ represents the
foreground heat map predicted by the network; $ M $ is the target heat map, $ W_U $ is the weight
parameter of the convolutional layers; $ \beta $ is a tradeoff parameter for weight decay.

[0033] Note that the sel-CNN for selecting features and the SNet and GNet for
localization are different in CNN structures. The sel-CNN architecture is very simple to
avoid using noisy feature maps to over fit the objective function, whereas the SNet and GNet
are more complex. Since the noisy feature maps have been discarded by the feature map
selection, more complex model facilitates more accurate tracking.

[0034] For a new frame, a region of interest (ROI) centered at the last target location
containing both target and background context is cropped and propagated through the fully
convolutional network such that two foreground heat maps are generated by the GNet and
the SNet, respectively, such that two foreground heat maps are generated, by the GNet and
the SNet, from the first selected feature maps from the higher layer of the CNN and the
selected second feature maps from the lower layer of the first CNN, respectively, and target
localization is performed independently based on the two heat maps, as will be further
discussed below.
In one embodiment of the present application, the target localization is first performed on the heat map produced by GNet. Denote the target location as $\hat{X} = (x, y, \sigma)$, where $x$, $y$ and $\sigma$ represent the center coordinates and scale of the target bounding box, respectively. Given the target location $\hat{X}^{-1}$ in the last frame, it assumes the locations of target candidates in the current frame are subject to a Gaussian distribution

$$p(X' | \hat{X}^{-1}) = \mathcal{N}(X' ; \hat{X}^{-1}, \Sigma)$$

(6)

where $\Sigma$ is a diagonal covariance matrix that indicates the variances of the location parameters. The confidence of the $i$-th candidate is computed as the summation of all the heatmap values within the candidate region.

$$conf_i = \sum_{j \in R_i} \hat{M}_G(j)$$

(7)

where $\hat{M}_G$ denotes the heat map generated by GNet; $R_i$ is the region of the $i$-th target candidate according to its location parameter $X_j(9)$; $j$ denotes the coordinate index. The candidate with the highest confidence is predicted as the target by GNet.

**The target localization unit 30**

[0036] The target localization unit 30, based on the two predicated heat maps, estimates the target location in the current frame via a distracter detection mechanism.

[0037] As discussed, GNet based on the higher layer (for example, conv5-3 layer) captures semantic features and is highly invariant to intra class variation. Hence, the foreground heat map generated by GNet highlights both the target and background distracters with similar appearances.

[0038] To prevent the tracker from drifting to background, the present application further promotes a distracter detection to determine the final target location. Denote the
target location predicted by GNet as $\hat{X}_G$, the corresponding target region in the heat map as $R_G$. The probability of background distracter is evaluated by the proportion between the confidence values outside and inside the target region

$$P_d = \frac{\sum_{j \in \hat{M}_G \setminus R_G} \hat{M}_G(j)}{\sum_{k \in R_G} \hat{M}_G(k)}$$

(8)

where $\hat{M}_G - R_G$ represents the background region on the predicted heat map $\hat{M}_G$. When the proportion $P_d$ is less than a predefined threshold (0.2 in all the experiments), it assume no background distracter appears and use the target location predicted by GNet as the final result. Otherwise, the same target localization procedure described above is performed on the heat map predicted by the SNet, and the candidate with the highest confidence is labeled as the target.

[0039] Returning to the target localization unit 30 according to one embodiment of the present application, it is configured to sample a set of target candidate regions according to a Gaussian distribution centered at the predicted target location in the last frame of the video sequences; predict a best target candidate in the current frame based on the target heat map estimated by the GNet, wherein the target confidence of each candidate is computed by a summation of heat map values within each of the candidate regions, and the candidate with the highest confidence is selected as the best target candidate; and compare the heat map values within a background region with those in the best candidate region to detect a distracter. If no distracter is detected, the best target location predicted using the heat map from the GNet as a final target location in the current frames, otherwise, a target localization using the specific heat map from the SNet will be utilized to predict the final target location.

The update unit 40

[0040] The update unit 40 updates the heat map prediction unit using previous tracking results in an online fashion to adapt to target appearance changes.
In order to avoid the background noise introduced by online update, the present application is proposed to fix GNet and only update SNet after the initialization in the first frame. SNet is updated following two different rules: the adaptation rule and the discrimination rule, which aim to adapt SNet to target appearance variation and improve the discriminative power for foreground and background, respectively. According to the adaptation rule, we fine-tune SNet every predetermined number of (for example, 20) frames using the most confident (for example, 10) tracking result within these predetermined number of frames. Based on the discrimination rule, when distracters are detected, SNet is further updated using the tracking results in the first frame and the current frame by minimizing the following objective function:

$$\min \beta \|W_s\| + \sum_{x,y} \left[ \left( \hat{M}_s^t(x,y) - M^t(x,y) \right)^2 + \left[ 1 - \Phi'(x,y) \right] \left[ \hat{M}_s^t(x,y) - M'(x,y) \right]^2 \right]$$  \hspace{1cm} (9)$$

where $W_s$ denotes the convolutional weight of SNet; $(x', y')$ are spatial coordinates; $\hat{M}_s^t$ and $M'$ represent the heat map for the $t$-th frame predicted by SNet and the heat map generated according to the predicted target location (a 2-dimensional Gaussian centered at the target location), respectively. The foreground mask $\Phi'$ indicates the predicted target bounding box, i.e., $\Phi'(x', y') = 1$ if the location $(x, y)$ belongs to the target region, and $\Phi'(x, y) = 0$, otherwise.

The second term in formula (9) corresponds to loss for locating the target object in the first frame. When distracters appear or the target undergoes severe occlusion, the estimated target region is not reliable for learning the appearance. Therefore, it chooses a conservative scheme in adding the first frame for supervising the learning so that the learned model still has the appearance in the first frame considered. The estimated target region is not used for updating the model because it is not reliable. Therefore, the third term in formula (9) removes the loss for the target region and only considers loss for the background.
region in the t-th frame. It enforces the model to put more efforts on assigning the background distracters as background. The combination of the second term and the third term in formula (9) can help SNet to better separate the target from background and alleviate the model degradation caused by occlusion or distracters.

[0043] Hereinafter, a method 3000 for object tracking according to one embodiment of the present application will be discussed in reference to Fig. 3. As shown, at step s31, a first frame of a video sequences is fed forward through a first CNN (convolutional network) to obtain a plurality of first feature maps for objects to be tracked in a higher layer of the CNN and a plurality of second feature maps for the objects in a lower layer of the first CNN. In particular, a region of interest (ROI) in a first frame of a video sequences will be determined (the ROI centered at the target location includes both the target region and background context) and the determined ROI is then fed forward through the first CNN to obtain a plurality of first feature maps for objects to be tracked in a higher layer of the CNN and a plurality of second feature maps for the objects in a lower layer of the first CNN.

[0044] At step S32, the method trains a second CNN and a third CNN (sel-CNNs) with the first feature maps and the second feature maps, respectively, to select a plurality of feature maps from the first and second feature maps, respectively. In particular, the first CNN is pre-trained on image classification task such that the first feature maps include more semantic features to determine a category for objects to be tracked in the video sequences, while the second feature maps carry more discriminative information to separate the objects from distracters with similar appearance.

[0045] Fig. 4 is a schematic diagram illustrating a flow chart for the training S32 according to one embodiment of the present application. As shown, the training further comprises a step S321 of initializing the two sel-CNNs with the first feature maps and the second feature maps, respectively, with the two features maps respectively to output a
predicted location for the objects in each sel-CNNs. At step S322, the predicted location is compared with a ground truth target location for the objects to obtain a prediction error, and step s323, it back-propagates the error through the sel-CNNs until the obtained error is less than a threshold.

[0046] At step S324, it determines a significance of each of the feature maps outputted from each of the two sel-CNNs, and all the feature maps are ranked in a descending order according to their significance values, wherein top ranked K feature maps from both higher and lower layers will be selected, wherein K is an integer greater than or equal to 1.

[0047] At step S325, at the online tracking stage for following frames, the feature maps are extracted and their corresponding K features maps are selected from the extracted maps as the first and the second feature maps.

[0048] Returning to Fig. 3, the method 3000 further comprises a step S33 of predicting, based on the first and the second feature maps, two target heat maps indicating the target location in a current frame, respectively, by the trained second CNN and a third CNN. Fig. 5 is a schematic diagram illustrating a flow chart for the step S33 of predicting according to one embodiment of the present application. As shown, in step S501, GNet and SNet are initialized by inputting the two features maps into the two CNNs respectively to output a heat map for the objects in each sel-CNNs; comparing the heat map with a ground truth heat map for the objects to obtain a prediction error; and back-propagating the error through the sel-CNNs until the obtained error is less than a threshold, and target heat maps for the first frame will be obtained. Then, at step S502, the initialized GNet and SNet estimate the target heat maps independently for following each frame, wherein a region of interest (ROI) centered at the last target location contains both target and background context and is cropped and propagated the CNN such that two foreground heat maps are generated, by the GNet and the SNet, from the selected first feature maps from the higher layer of the CNN.
and the selected second feature maps from the lower layer of the first CNN, respectively, and target localization is performed independently based on the two heat maps. In one embodiment of the present application, each of the GNet and SNet consists of a first convolutional layer and a second convolutional layer nonlinearly connected to the first layer, wherein the first convolutional layer has a larger convolutional kernel of size and the second convolutional layer has kernels with a relative small size.

[0049] Returning to Fig. 3 again, the method 3000 further comprises a step S34, in which, a target location for the object will be estimated based on the two predicated heat maps in the current frame. In one embodiment of the present application, the step S34 may comprise as shown in Fig. 6:

step S341: sampling a set of target candidate regions according to a Gaussian distribution centered at the predicted target location in the last frame of the video sequences;

step S342: predicting a best target candidate in the current frame based on the target heat map estimated by the GNet, wherein the target confidence of each candidate is computed by a summation of heat map values within each of the candidate region, and the candidate with the highest confidence is selected as the best target candidate;

step S343: compare the heat map values within a background region with those in the best candidate region to detect a distracter;

step S344: it is determined if the distracter is detected. If no distracter is detected, at step S345 the best target location predicted using the heat map from the GNet is determined as a final target location in the current frames at step S345, otherwise, a target localization using the specific heat map from the SNet will be utilized to predict the final target location at step S346.

[0050] Fig. 7 illustrates a system 700 for object tracking consistent with an embodiment of the present application. In this embodiment, the process as shown in Fig. 3 can be implemented by software in the system 700. Referring to Fig. 7, the system 700
comprises a memory 71 that stores executable components and a processor 72, electrically 
coupled to the memory 71 to execute the executable components to perform operations of the 
system 700. The executable components may comprise: a feature map selection 
component73 configured to determine a region of interest (ROI) in a first frame of a video 
sequences, the ROI including pixels for objects to be tracked, and feed the determined ROI 
forward through a first CNN (convolutional network), and obtain a plurality of first feature 
maps in a higher layer of the CNN and a plurality of second feature maps in a lower layer of 
the first CNN; a heat map prediction component74 configured to predict, based on the first 
and the second feature maps, two target heat maps indicating a target location in the current 
frame, respectively; and a target localization component 75 configured to, based on the two 
predicated heat maps, estimate a final target location for the ROI in the current frame.

[0051] As discussed in the above, the first CNN may be pre-trained on image 
classification task such that the first feature maps include more semantic features to 
determine a category for objects to be tracked in the video sequences, while the second 
feature maps carry more discriminative information to separate the objects from distracters 
with similar appearance.

[0052] The same as the feature map selection unit as discussed in the above, the feature 
map selection component is configured with two convolutional neural networks (i.e., 
sel-CNNs), and wherein the feature map selection unit is configured to, initialize the two 
sel-CNNs with the first feature maps and the second feature maps, respectively, by inputting 
the two features maps into the two CNNs respectively to output a heat map for the objects in 
each sel-CNNs; comparing the heat map with a ground truth heat map for the objects to 
obtain a prediction error; and back-propagating the error through the sel-CNNs until the 
obtained error is less than a threshold; determine a significance for each of those in the first 
and second feature maps according to the two trained sel-CNNs; rank those in the first and 
second feature maps independently in a descending order according to their significance
values; and select top ranked K feature maps from both higher and lower layers, wherein K is an integer greater than or equal to 1; wherein at an online tracking stage for following frames, the first and second feature maps are extracted from the higher and lower layers of the first CNN respectively and their corresponding K features maps are selected and serve as said selected first and second feature maps.

[0053] In one embodiment, the heat map prediction component is configured with a general network and a specific network, wherein the general neural network (GNet) captures the category information of the target object based on top of the selected feature maps from the first feature maps; the specific neural network (SNet) discriminates the target object from background with similar appearance based on top of the selected second feature maps of the lower layer. The GNet and the SNet are initialized in the first frame to perform foreground heat map regression for the target object; for a new frame, a region of interest (ROI) centered at a last target location containing both target and background context is cropped and propagated through the fully convolutional network; and two foreground heat maps are generated by the GNet and the SNet, respectively.

[0054] Both GNet and SNet are initialized by following steps: feeding the selected first and second feature maps of the first frame through the GNet and SNet respectively to predict two target heat maps; comparing the predicted heat maps with a ground truth heat map to obtain prediction errors; back-propagating the errors through the GNet and SNet until the obtained errors are less than a threshold; and wherein the ground truth heat map is distributed in accordance with a 2-dimensional Gaussian distribution centered at the ground truth target location with variance proportional to a target size of the objects.

[0055] As discussed in reference with the target localization unit, the target localization component is configured to, sample a set of target candidate regions according to a Gaussian distribution centered at the predicted target location in the last frame of the video sequences;
predict a best target candidate in the current frame based on the target heat map estimated by the GNet, wherein the target confidence of each candidate is computed by a summation of heat map values within each of the candidate regions, and the candidate with the highest confidence is selected as the best target candidate; and compare the heat map values within a background region with those in the best candidate region to detect a distracter; if no distracter is detected, the best target location predicted using the heat map from the GNet is determined as a final target location in the current frames, otherwise, a target localization using the specific heat map from the SNet will be utilized to predict the final target location.

[0056] The first CNN consists of a dropout layer followed by a convolutional layer without any nonlinear transformation. Each of the GNet and SNet consists of a first convolutional layer and a second convolutional layer nonlinearly connected to the first layer, wherein the first convolutional layer has a larger convolutional kernel of size and the second convolutional layer has kernels with a relative small size.

[0057] As shown, the system 700 may further comprise an update component 76 configured to update the heat map prediction unit using previous tracking locations in an online fashion to adapt to target appearance changes.

[0058] As will be appreciated by one skilled in the art, the present invention may be embodied as a system, method or computer program product. Accordingly, the present invention may take the form of an entirely hardware embodiment and hardware aspects that may all generally be referred to herein as a "unit", "circuit," "module" or "system." Much of the inventive functionality and many of the inventive principles when implemented, are best supported with or integrated circuits (ICs), such as a digital signal processor and software therefore or application specific ICs. It is expected that one of ordinary skill, notwithstanding possibly significant effort and many design choices motivated by, for example, available time, current technology, and economic considerations, when guided by the concepts and principles
disclosed herein will be readily capable of generating ICs with minimal experimentation. Therefore, in the interest of brevity and minimization of any risk of obscuring the principles and concepts according to the present invention, further discussion of such software and ICs, if any, will be limited to the essentials with respect to the principles and concepts used by the preferred embodiments.

[0059] Although the preferred examples of the present invention have been described, those skilled in the art can make variations or modifications to these examples upon knowing the basic inventive concept. The appended claims are intended to be considered as comprising the preferred examples and all the variations or modifications fell into the scope of the present invention.

[0060] Obviously, those skilled in the art can make variations or modifications to the present invention without departing the spirit and scope of the present invention. As such, if these variations or modifications belong to the scope of the claims and equivalent technique, they may also fall into the scope of the present invention.
What is claimed is:

1. A method for object tracking, comprising:
   determining a region of interest (ROI) in a first frame of a video sequences, wherein the ROI is centered at a ground truth target location for objects to be tracked;
   feeding the determined ROI forward through a first CNN (convolutional network) to obtain a plurality of first feature maps in a higher layer of the CNN and a plurality of second feature maps in a lower layer of the first CNN, wherein the first CNN is pre-trained on an image classification task such that the first feature maps include more semantic features to determine a category for objects to be tracked in the video sequences, while the second feature maps carry more discriminative information to separate the objects from distracters with similar appearance;
   selecting a plurality of feature maps from the first and second feature maps, respectively by training a second CNN and a third CNN (sel-CNNs) with the first feature maps and the second feature maps, respectively;
   predicting, based on the selected first and second feature maps, two target heat maps indicating a target location for said objects in the current frame, respectively; and
   estimating, based on the two predicated target heat maps, a final target location for the object in the current frame.

2. The method according to claim 1, wherein the training further comprises:
   initializing the two sel-CNNs with the first feature maps and the second feature maps, respectively, with the two features maps respectively to output a heat map for the objects in each of the sel-CNNs;
   comparing the heat map with a ground truth heat map for the objects to obtain a prediction error for each of the sel-CNNs; and
   back-propagating the error through each of the sel-CNNs until the obtained error is less than a threshold.
3. The method according to claim 2, wherein the training further comprises:
determining a significance for each of those in the first and second feature maps according to
the two trained sel-CNNs;
ranking those in the first and second feature maps independently in a descending order
according to their significance values; and
selecting top ranked K feature maps from both higher and lower layers, wherein K is an
integer greater than or equal to 1;
wherein at an online tracking stage for following frames, the first and second feature maps
are extracted from the higher and lower layers of the first CNN respectively and their
corresponding K features maps are selected and serve as said selected first and second
feature maps.

4. The method according to claim 1, wherein the predicting comprises:
initializing a GNet and a SNet and obtaining target heat maps for the first frame;
estimating, by the initialized GNet and SNet, the target heat maps independently for
following each frame,
wherein the ROI centered at a last target location contains both target and background
context and is cropped and propagated through the first CNN to obtain the first and second
feature maps, and the selected first and second feature maps are propagated through the GNet
and the SNet, respectively; and
wherein two foreground heat maps are generated by the GNet and the SNet, respectively, and
a target localization prediction is performed independently based on the two foreground heat
maps.

5. The method according to claim 4, wherein both GNet and SNet are initialized by:
feeding the selected first and second feature maps of the first frame through the GNet and
SNet respectively to predict two target heat maps;
comparing the predicted heat maps with a ground truth heat map to obtain prediction errors; 
back-propagating the errors through the GNet and SNet until the obtained errors are less than 
a threshold; and 
wherein the ground truth heat map is distributed in accordance with a 2-dimensional 
Gaussian distribution centered at the ground truth target location with variance proportional 
to a target size of the objects.

6. The method according to claim 5, wherein the estimating further comprises:
sampling a set of target candidate regions according to a Gaussian distribution centered at the 
predicted target location in a last frame of the video sequences; 
predicting a best target candidate in the current frame based on the target heat map estimated 
by the GNet, wherein the target confidence of each candidate is computed by a summation of 
heat map values within each of the candidate regions, and the candidate with the highest 
confidence is selected as the best target candidate, 
comparing the heat map values within a background region with those in the best candidate 
region to detect a distracter; 
if no distracter is detected, the best target location predicted using the heat map from the 
GNet is determined as a final target location in the current frames, otherwise, a target 
localization using the specific heat map from the SNet will be utilized to predict the final 
target location.

7. The method according to claim 1, wherein each of the GNet and SNet consists of a first 
convolutional layer and a second convolutional layer nonlinearly connected to the first 
convolutional layer, 
wherein the first convolutional layer has a larger convolutional kernels of size and the second 
convolutional layer has kernels with a relative small size.

8. The method according to any one of claims 4-7, further comprising:
updating the SNet with previous tracking location for the objects in an online fashion to adapt to target appearance changes.

9. A system for object tracking, comprising:
   a feature map selection unit configured to determine a region of interest (ROI) in a first frame of a video sequences, wherein the ROI is centered at a ground truth target location for objects to be tracked, and feed the determined ROI forward through a first CNN (convolutional network) to obtain a plurality of first feature maps in a higher layer of the CNN and a plurality of second feature maps in a lower layer of the first CNN;
   a heat map prediction unit electrically communicated with the feature map selection unit and configured to predict, based on the first and the second feature maps, two target heat maps indicating a target location in a current frame, respectively; and
   a target localization unit configured to, based on the two predicated heat maps, estimate a final target location for the ROI in the current frame,
   wherein the first CNN is pre-trained on an image classification task such that the first feature maps include more semantic features to determine a category for objects to be tracked in the video sequences, while the second feature maps carry more discriminative information to separate the objects from distracters with similar appearance.

10. The system according to claim 9, wherein the feature map selection unit is configured with two convolutional neural networks (sel-CNNs), and wherein the feature map selection unit is configured to,
   initialize the two sel-CNNs with the first feature maps and the second feature maps, respectively, by inputting the two features maps into the two CNNs respectively to output a heat map for the objects in each of the sel-CNNs; comparing the heat map with a ground truth heat map for the objects to obtain a prediction error in each of the sel-CNNs; and back-propagating the error through the sel-CNNs until the obtained error is less than a threshold;
determine a significance for each of those in the first and second feature maps according to the two trained sel-CNNs;
rank those in the first and second feature maps independently in a descending order according to their significance values; and
select top ranked K feature maps from both higher and lower layers, wherein K is an integer greater than or equal to 1;
wherein at an online tracking stage for following frames, the first and second feature maps are extracted from the higher and lower layers of the first CNN respectively and their corresponding K features maps are selected and serve as the selected first and second feature maps.

11. The system according to claim 10, wherein the heat map prediction unit is configured with a general network (GNet) and a specific network (SNet), wherein the general network captures the category information of the target object based on top of the selected feature maps from the first feature maps; and the specific network discriminates the target object from background with similar appearance based on top of the selected second feature maps of the lower layer.

12. The system according to claim 11, wherein, the GNet and the SNet are initialized in the first frame to perform foreground heat map regression for the target object; for a new frame, the ROI is centered at a last target location containing both target and background context and is cropped and propagated through the first CNN such that two foreground heat maps are generated, by the GNet and the SNet, from the selected first feature maps from the higher layer of the first CNN and the selected second feature maps from the lower layer of the first CNN, respectively.

13. The system according to claim 11, wherein both GNet and SNet are initialized by: feeding the selected first and second feature maps of the first frame through the GNet and the
SNet respectively to predict two target heat maps;
comparing the predicted heat maps with a ground truth heat map to obtain prediction errors;
back-propagating the errors through the GNet and SNet until the obtained errors are less than a threshold; and
wherein the ground truth heat map is a 2-dimensional Gaussian distribution centered at the ground truth target location with variance proportional to a target size of the objects.

14. The system according to claim 12, wherein the target localization unit is configured to,
sample a set of target candidate regions according to a Gaussian distribution centered at the predicted target location in a last frame of the video sequences;
predict a best target candidate in the current frame based on the target heat map estimated by the GNet, wherein the target confidence of each candidate is computed by a summation of heat map values within each of the candidate region, and the candidate with the highest confidence is selected as the best target candidate; and
compare the heat map values within a background region with those in the best candidate region to detect a distracter;
if no distracter is detected, the best target location predicted using the heat map from the GNet is determined as a final target location in the current frames, otherwise, a target localization using the specific heat map from the SNet will be utilized to predict the final target location.

15. The system according to claim 9, wherein the sel-CNNs consist of a dropout layer followed by a convolutional layer without any nonlinear transformation.

16. The system according to claim 11, wherein each of the GNet and the SNet consists of a first convolutional layer and a second convolutional layer nonlinearly connected to the first convolutional layer,
wherein the first convolutional layer has a larger convolutional kernels of size and the second
convolutional layer has kernels with a relative small size.

17. The system according to any one of claims 9-16, further comprising:
an update unit configured to update the heat map prediction unit using previous tracking locations in an online fashion to adapt to target appearance changes.

18. A system for object tracking, comprising:
a memory that stores executable components; and
a processor that executes the executable components to perform operations of the system, the executable components comprise:
a feature map selection component configured to determine a region of interest (ROI) in a first frame of a video sequences, wherein the ROI is centered at aground truth target location for objects to be tracked, and feed the determined ROI forward through a first CNN (convolutional network) to obtain a plurality of first feature maps in a higher layer of the CNN and a plurality of second feature maps in a lower layer of the first CNN;
a heat map prediction component configured to predict, based on the first and the second feature maps, two target heat maps indicating a target location in the current frame, respectively; and
a target localization component configured to, based on the two predicated heat maps, estimate a final target location for the ROI in the current frame.

19. The system according to claim 18, wherein the first CNN is pre-trained on image classification task such that the first feature maps include more semantic features to determine a category for objects to be tracked in the video sequences, while the second feature maps carry more discriminative information to separate the objects from distracters with similar appearance.

20. The system according to claim 19, wherein the feature map selection component is
configured with two convolutional neural networks (i.e., sel-CNNs), and
wherein the feature map selection unit is configured to,
initialize the two sel-CNNs with the first feature maps and the second feature maps, respectively, by inputting the two features maps into the two CNNs respectively to output a heat map for the objects in each of the sel-CNNs; comparing the heat map with a ground truth heat map for the objects to obtain a prediction error in each of the sel-CNN; and back-propagating the error through the sel-CNNs until the obtained error is less than a threshold;
determine a significance for each of those in the first and second feature maps according to the two trained sel-CNNs;
rank those in the first and second feature maps independently in a descending order according to their significance values; and
select top ranked K feature maps from both higher and lower layers, wherein K is an integer greater than or equal to 1;
wherein at an online tracking stage for following frames, the first and second feature maps are extracted from the higher and lower layers of the first CNN respectively and their corresponding K features maps are selected and serve as said selected first and second feature maps.

21. The system according to claim 18, wherein the heat map prediction component is configured with a general network (GNet) and a specific network(SNet),
wherein the GNet captures the category information of the target object based on top of the selected feature maps from the first feature maps; and
the SNet discriminates the target object from background with similar appearance based on top of the selected second feature maps of the lower layer.

22. The system according to claim 21, wherein, the GNet and the SNet are initialized in the first frame to perform foreground heat map regression for the target object;
for a new frame, the ROI centered at a last target location contains both target and background context and is cropped and propagated through the first CNN such that two foreground heat maps are generated, by the GNet and the SNet, from the selected first feature maps from the higher layer of the first CNN and the selected second feature maps from the lower layer of the first CNN, respectively.

23. The system according to claim 22, wherein both the GNet and the SNet are initialized by:

- feeding the selected first and second feature maps of the first frame through the GNet and the SNet respectively to predict two target heat maps;
- comparing the predicted heat maps with a ground truth heat map to obtain prediction errors;
- back-propagating the errors through the GNet and the SNet until the obtained errors are less than a threshold; and

wherein the ground truth heat map is distributed in accordance with a 2-dimensional Gaussian centered at the ground truth target location with variance proportional to a target size of the objects.

24. The system according to claim 22, wherein the target localization component is configured to,

- sample a set of target candidate regions according to a Gaussian distribution centered at the predicted target location in a last frame of the video sequences;
- predict a best target candidate in the current frame based on the target heat map estimated by the GNet, wherein the target confidence of each candidate is computed by a summation of heat map values within each of the candidate regions, and the candidate with the highest confidence is selected as the best target candidate; and
- compare the heat map values within a background region with those in the best candidate region to detect a distracter;

if no distracter is detected, the best target location predicted using the heat map from the GNet is determined as a final target location in the current frames, otherwise, a target
localization using the specific heat map from the SNet will be utilized to predict the final target location.

25. The system according to claim 18, wherein the sel-CNNs consist of a dropout layer followed by a convolutional layer without any nonlinear transformation.

26. The system according to claim 18, wherein each of the GNet and SNet consists of a first convolutional layer and a second convolutional layer nonlinearly connected to the first convolutional layer, wherein the first convolutional layer has a larger convolutional kernels of size and the second convolutional layer has kernels with a relative small size.

27. The system according to any one of claims 18-26, further comprising: an update component configured to update the heat map prediction unit using previous tracking locations in an online fashion to adapt to target appearance changes.
Fig. 2

FEATURE MAP SELECTION UNIT

HEAT MAP PREDICTION UNIT

MODEL UPDATE UNIT

TARGET LOCALIZATION UNIT

Target Location
3000

S31

Obtain first feature maps and second feature maps for the objects from a first CNN

S32

Training second and third CNNs to select features from the first and the second feature maps

S33

Predicting two target heat maps indicating target location in a current frame

S34

Estimating a final target location for the object in the current frame

Fig. 3
S32

S321
INITIALIZE THE TWO SEL-CNNS

S322
OBTAIN A PREDICTION ERROR

S323
BACK-PROPAGATES THE ERROR UNTIL CONVERGENCE

S324
SELECT TOP RANKED K FEATURE MAPS IN FIRST FRAME

S25
FOR FOLLOWING FRAMES, FEATURE MAPS ARE EXTRACTED AND THEIR CORRESPONDING K FEATURES MAPS ARE SELECTED

Fig. 4
S33

INITIALIZING GNET AND SNET AND OBTAINING TARGET HEAT MAPS FOR THE FIRST FRAME

S331

IN THE FOLLOWING EACH FRAME, THE TRAINED GNET AND SNET TAKES THE SELECTED FEATURE MAPS AS INPUT AND ESTIMATE THE TARGET HEAT MAPS INDEPENDENTLY

S332

Fig. 5
S34

SAMPLING A SET OF TARGET CANDIDATE REGIONS

S341

S342

PREDICTING A BEST TARGET CANDIDATE IN THE CURRENT FRAME BASED ON THE TARGET HEAP MAP ESTIMATED BY THE GNET

S343

COMPARING THE HEAT MAP VALUES WITHIN A BACKGROUND REGION WITH THOSE IN THE BEST CANDIDATE REGION TO DETECT A DISTRACTER

S344

DISTRACTER?

S345

THE TARGET LOCATION PREDICTED USING THE GENERIC HEAT MAP WILL BE OUTPUT AS THE FINAL TARGET LOCATION

S346

PERFORMS TARGET LOCALIZATION USING THE SPECIFIC HEAT MAP FROM THE SNET

Fig. 6
Fig. 7
INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER
G06T 7/00(2006.01)i
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED
Minimum documentation searched (classification system followed by classification symbols)
G06T; H04N
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
WPLEPDOC,CNPAT,CNKI:region,interest,ROI,convolutional network,CNN,feature map,layer,semantic,classification, category,target,object,track,location

C. DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>CN 103260081 A (CHINA MOBILE COMMUNICATION CORPORATION) 21 August 2013 (2013-08-21) the whole document</td>
<td>1-27</td>
</tr>
<tr>
<td>A</td>
<td>US 2010110183 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 06 May 2010 (2010-05-06) the whole document</td>
<td>1-27</td>
</tr>
</tbody>
</table>

Further documents are listed in the continuation of Box C.  
* Special categories of cited documents:
  "A" document defining the general state of the art which is not considered to be of particular relevance
  "E" earlier application or patent but published on or after the international filing date
  "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
  "O" document referring to an oral disclosure, use, exhibition or other means
  "P" document published prior to the international filing date but later than the priority date claimed
  "Y" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
  "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
  "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
  "&" document member of the same patent family

Date of the actual completion of the international search 01 March 2016
Date of mailing of the international search report 27 April 2016

Name and mailing address of the ISA/CN

STATE INTELLECTUAL PROPERTY OFFICE OF THE P.R.CHINA
6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088, China

Facsimile No. (86-10)62019451

Authorized officer

SUN, Zhifei

Telephone No. (86-10)62413325

Form PCT/ISA/210 (second sheet) (July 2009)
<table>
<thead>
<tr>
<th>Patent document cited in search report</th>
<th>Publication date (day/month/year)</th>
<th>Patent family member(s)</th>
<th>Publication date (day/month/year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN 104680508 A</td>
<td>03 June 2015</td>
<td>WO 2015078185 A1</td>
<td>04 June 2015</td>
</tr>
<tr>
<td>CN 103260081 A</td>
<td>21 August 2013</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>US 2010110183 A1</td>
<td>06 May 2010</td>
<td>None</td>
<td></td>
</tr>
</tbody>
</table>

Form PCT/ISA/210 (patent family annex) (July 2009)