202329121

(19)中華民國智慧財產局

(12)發明說明書公開本 (11)公開編號：TW 202329121 A

(43)公開日： 中華民國 112 (2023) 年 07 月 16 日

(21)申請案號：111120620 (22)申請日： 中華民國 111 (2022) 年 06 月 02 日

(51)Int. Cl.： *G11C16/10 (2006.01)*

(30)優先權：2022/01/11 美國 63/298,612

(71)申請人：旺宏電子股份有限公司 (中華民國) MACRONIX INTERNATIONAL CO., LTD. (TW)

新竹縣科學工業園區力行路 16 號

(72)發明人：王韋程 WANG, WEI-CHEN (TW) ；胡瀚文 HU, HAN-WEN (TW) ；李永駿 LI, YUNG-CHUN (TW) ；王淮慕 WANG, HUAI-MU (TW) ；何建忠 HO, CHIEN-CHUNG (TW) ；張原豪 CHANG, YUAN-HAO (TW) ；郭大維 KUO, TEI-WEI (TW)

(74)代理人：祁明輝；林素華

申請實體審查：有 申請專利範圍項數：10 項 圖式數：5 共 38 頁

(54)名稱

記憶體裝置及其操作方法

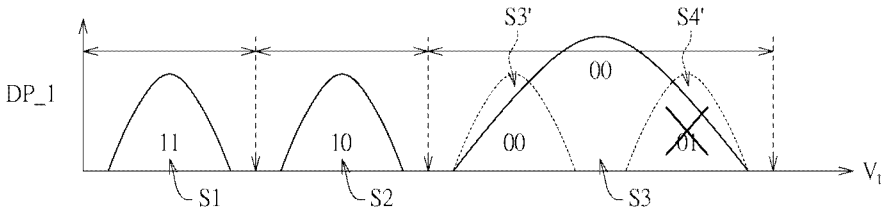
(57)摘要

係揭示一種記憶體裝置，包括記憶體陣列，用於儲存多個向量資料。各向量資料具有高有效位元(MSB)向量與低有效位元(LSB)向量。記憶體陣列包括多個記憶單元，各記憶單元具有第一位元及第二位元。第一位元用於儲存各向量資料的 MSB 向量，第二位元用於儲存各向量資料的 LSB 向量。各該向量資料對應的一位元線執行一次的位元線設定，並根據該位元線讀取各該向量資料的該 MSB 向量與該 LSB 向量。各記憶單元的臨界電壓分布區分為 N 個狀態，N 為正整數且 N 小於 2 的 2 次方，各記憶單元儲存的等效位元數小於 2。

A memory device is disclosed, which includes a memory array for storing a plurality of vector data. Each of the vector data has an MSB vector and a LSB vector. The memory array includes a plurality of memory units, each of the memory units has a first bit and a second bit. The first bit is used to store the MSB vector of each vector data, and the second bit is used to store the LSB vector of each vector data. A bit line corresponding to each vector data executes one time of bit-line-setup, and reads the MSB vector and the LSB vector of each vector data according to the bit line. The threshold voltage distribution of each memory unit is divided into N states, where N is a positive integer and N is less than 2 to the power of 2, and the effective bit number stored by each memory unit is less than 2.

指定代表圖：

TW 202329121 A



第 4A 圖

符號簡單說明：

S1~S3,S3',S4':狀態

DP_1:第一資料樣式

Vt:臨界電壓

專利案號：111120620
202329121

【發明摘要】

【中文發明名稱】記憶體裝置及其操作方法

【英文發明名稱】MEMORY DEVICE AND OPERATING METHOD THEREOF

【中文】

係揭示一種記憶體裝置，包括記憶體陣列，用於儲存多個向量資料。各向量資料具有高有效位元(MSB)向量與低有效位元(LSB)向量。記憶體陣列包括多個記憶單元，各記憶單元具有第一位元及第二位元。第一位元用於儲存各向量資料的MSB向量，第二位元用於儲存各向量資料的LSB向量。各該向量資料對應的一位元線執行一次的位元線設定，並根據該位元線讀取各該向量資料的該MSB向量與該LSB向量。各記憶單元的臨界電壓分布區分為N個狀態，N為正整數且N小於2的2次方，各記憶單元儲存的等效位元數小於2。

【英文】

A memory device is disclosed, which includes a memory array for storing a plurality of vector data. Each of the vector data has an MSB vector and a LSB vector. The memory array includes a plurality of memory units, each of the memory units has a first bit and a second bit. The first bit is used to store the MSB vector of each vector data, and the second bit is used to store the LSB vector of each vector data. A bit line corresponding to each vector data executes one time of bit-line-setup, and reads the MSB vector and the LSB vector of each vector data according to the bit line. The threshold voltage distribution of each

memory unit is divided into N states, where N is a positive integer and N is less than 2 to the power of 2, and the effective bit number stored by each memory unit is less than 2.

【指定代表圖】第4A圖

【代表圖之符號簡單說明】

S1~S3,S3',S4':狀態

DP_1:第一資料樣式

Vt:臨界電壓

【特徵化學式】無

【發明說明書】

【中文發明名稱】記憶體裝置及其操作方法

【英文發明名稱】MEMORY DEVICE AND OPERATING METHOD THEREOF

【技術領域】

【0001】本揭示係關於一種半導體裝置及其操作方法，特別有關於一種記憶體裝置及記憶體內運算的方法。

【先前技術】

【0002】人工智慧的演算法迅速發展，研發人員尋求適於執行人工智慧的演算法的硬體裝置。半導體的記憶體裝置，例如NAND型的記憶體陣列，可藉由記憶體內運算(in-memory computing，IMC)以執行人工智慧的運算。當執行人工智慧常用的乘積累加(Multiply-and-Accumulate，MAC)運算時，記憶體陣列的每個記憶單元可根據乘積運算結果輸出電流至位元線，並在同一條位元線上累加電流而得到累加運算結果。

【0003】在記憶體陣列中，以類比方式執行MAC運算可達到較快的運算速度。然而，電流分布的重疊(overlap)可能導致運算結果的誤判，而降低運算精確度。

【0004】另一方面，當讀取記憶體陣列儲存的資料時，可能需要執行多次的位元線設定(bit line setup)而耗費多次的設定時間(setup time)，因而降低運算速度。

【0005】此外，當記憶體陣列採用多階記憶單元(multi-level cell，MLC)(或稱為「二階記憶單元」)儲存資料時，記憶單元的臨界電壓分布具有較多數量的狀態，將導致讀取電壓的電壓間隔縮小，可能造成讀取操作的錯誤。

【0006】有鑑於上述技術問題，本技術領域之相關產業之技術人員係致力於改良記憶體單元的儲存機制及IMC運算的操作方法，以提升運算精準度及運算速度。

【發明內容】

【0007】本揭示之技術方案係提供以記憶體陣列執行數位方式的 MAC 運算、減少記憶單元的臨界電壓分布的狀態的數量、減少位元線設定的次數，以提升運算精準度及運算速度。

【0008】根據本揭示之一方面，提供一種記憶體裝置，包括記憶體陣列，用於儲存多個向量資料，各向量資料具有高有效位元(MSB)向量與低有效位元(LSB)向量，記憶體陣列包括多個記憶單元，各記憶單元具有第一位元及第二位元，第一位元用於儲存各向量資料的 MSB 向量，第二位元用於儲存各向量資料的 LSB 向量。各向量資料對應的位元線執行一次的位元線設定，並根據位元線讀取各向量資料的 MSB 向量與 LSB 向量。各記憶單元的臨界電壓分布區分為 N 個狀態，N 為正整數且 N 小於 2 的 2 次方，各記憶單元儲存的等效位元數小於 2。

【0009】根據本揭示之另一方面，提供一種記憶體裝置的操作方法，其中記憶體裝置包括記憶體陣列，記憶體陣列包括多個記憶單元，

各記憶單元具有第一位元及第二位元，操作方法包括以下步驟。儲存多個向量資料於記憶體陣列，各向量資料具有高有效位元(MSB)向量與低有效位元(LSB)向量。儲存各向量資料的 MSB 向量於第一位元。儲存各向量資料的 LSB 向量於第二位元。對於各該向量資料對應的一位元線執行一次的位元線設定。根據該位元線讀取各該向量資料的該 MSB 向量與該 LSB 向量。各記憶單元的臨界電壓分布區分為 N 個狀態，N 為正整數且 N 小於 2 的 2 次方，各記憶單元儲存的等效位元數小於 2。

【0010】透過閱讀以下圖式、詳細說明以及申請專利範圍，可見本揭示之其他方面以及優點。

【圖式簡單說明】

【0011】

第 1A 圖為本揭示一實施例的記憶體裝置的示意圖。

第 1B 圖為本揭示的記憶體裝置執行的 MAC 運算的流程圖。

第 2A 圖為常規的 2 位元 MLC 記憶單元的臨界電壓分布的狀態的示意圖。

第 2B 圖為本揭示一實施例的 1.5 位元 MLC 記憶單元的臨界電壓分布的狀態的示意圖。

第 3 圖為本揭示的記憶體裝置執行的 MAC 運算之另一實施例的流程圖。

第 4A~4D 圖為本揭示的不同實施例的 1.5 位元 MLC 記憶單元的臨界電壓分布的狀態的示意圖。

第 5A 圖為常規的 3 位元 TLC 記憶單元的臨界電壓分布的狀態的示意圖。

第 5B 圖為本揭示一實施例的 TLC 記憶單元的臨界電壓分布的狀態的示意圖。

第 5C~5I 圖為本揭示的不同實施例的 TLC 的臨界電壓分布的狀態的示意圖。

【實施方式】

【0012】本說明書的技術用語係參照本技術領域之習慣用語，如本說明書對部分用語有加以說明或定義，該部分用語之解釋係以本說明書之說明或定義為準。本揭露之各個實施例分別具有一或多個技術特徵。在可能實施的前提下，本技術領域具有通常知識者可選擇性地實施任一實施例中部分或全部的技術特徵，或者選擇性地將這些實施例中部分或全部的技術特徵加以組合。

【0013】第 1A 圖為本揭示一實施例的記憶體裝置 1000 的示意圖。請參見第 1A 圖，記憶體裝置 1000 包括記憶體陣列 100 及複數個資料鎖存器(data latch) 210、220 及 230。記憶體裝置 1000 可用於執行記憶體內運算(in-memory computation，IMC)，例如執行乘積累加(multiply-and-accumulate，MAC)運算。更特定而言，本揭示的記憶體裝置 1000 用於執行數位的 MAC 運算以提升運算精準度。資料鎖存器 210 可稱為「第一資料鎖存器」，資料鎖存器 220 可稱為「第二資料鎖存器」。並且，資料鎖存器 230 可稱為「共同資料鎖存器(common data latch，CDL)」，可

作為記憶體裝置 1000 的輸入/輸出介面。

【0014】記憶體陣列 100 包括複數條字元線 WL1~WLN，此些字元線 WL1~WLN 的每一者可儲存一個向量(vector)資料。當記憶體裝置 1000 用於執行人臉辨識演算法的 MAC 運算時，N 個人臉 F1~FN 的特徵向量(feature vector) q1~qN 可預先儲存於記憶體陣列 100 的字元線 WL1~WLN。例如，第 1 個人臉 F1 的特徵向量 q1 為 20 個位元的二進位(binary)資料「00110011001100001111」而儲存於字元線 WL1，第 2 個人臉 F2 的特徵向量 q2 為 20 個位元的資料「11001100110000001111」而儲存於字元線 WL2，依此類推。

【0015】並且，本揭示的記憶體陣列 100 之中的記憶單元為 1.5 位元之多階記憶單元(multi-level cell，MLC)。相較於常規的 2 位元之 MLC 記憶單元(以下簡稱為「2 位元 MLC」)，本揭示的 1.5 位元之 MLC 記憶單元(以下簡稱為「1.5 位元 MLC」)儲存的資料量較少，等效位元數為 1.5 位元。1.5 位元 MLC 的機制將詳述於後文。

【0016】資料鎖存器230(即，共同資料鎖存器(CDL))可暫存由記憶體裝置1000外部輸入的向量資料。當執行人臉辨識或人臉搜尋時，待辨識的人臉F0的特徵向量p0可由外部輸入至記憶體裝置1000而暫存於資料鎖存器230。特徵向量p0例如為20個位元的「11111111000000001111」。而後，特徵向量p0由資料鎖存器230轉移(transfer)至資料鎖存器210。

【0017】可在記憶體陣列 100 中執行「選擇位元線讀取操作(selective BL read)」從字元線 WL1~WLN 分別讀取特徵向量 q1~qN，選擇位元線讀取操作可具有乘積運算的功能。待辨識的特徵向量 p0 亦可稱為「輸入特徵向量 p0」，預先儲存的特徵向量 q1~qN 亦可稱為「權重特徵向量 qi」，i=1~N。選擇位元線讀取操作可達成輸入特徵向量 p0 與權重特徵向量 qi 的乘積運算。當特徵向量 p0 的位元與特徵向量 qi 的對應位元皆為「1」時，選擇位元線讀取操作可將特徵向量 qi 的該位元讀取為「1」。當特徵向量 p0 的位元為「0」或特徵向量 qi 的對應位元為「0」時，選擇位元線讀取操作可將特徵向量 qi 的該位元讀取為「0」。據此，選擇位元線讀取操作得到的輸出結果 qi'=p0*qi。輸出結果 qi'亦可稱為「輸出特徵向量 qi'」。例如，特徵向量 p0=「11111111000000001111」，第 1 個人臉 F1 的特徵向量 q1=「00110011001100001111」，選擇位元線讀取操作得到的輸出特徵向量 q1'=「00110011000000001111」。而後，輸出特徵向量 q1'可寫入資料鎖存器 220。

【0018】而後，可對於資料鎖存器 220 暫存的輸出特徵向量 q1'執行累加運算，以累加輸出特徵向量 q1'之中的「1」的位元。在累加運算的一種示例中，記憶體裝置 1000 可執行錯誤位元計數指令(fail-bit-count)以對於輸出特徵向量 q1'進行計數操作(counting operation)。以輸出特徵向量 q1'=「00110011000000001111」為例，其中的 8 個位元 B0~B3、B12、

B13、B16、B17 為「1」，輸出特徵向量 q1'的計數結果 C1 為十進位「8」，計數結果 C1 即為特徵向量 p0 與特徵向量 q1 的乘積累加結果 MAC1。而後，計數結果 C1 可轉移至資料鎖存器 230，並可將計數結果 C1 輸出至記憶體裝置 1000 的外部。

【0019】在人臉辨識演算法中，預先儲存的人臉 F1 與待辨識的人臉 F0 之間的相似度可表示為特徵向量 p0 與特徵向量 q1 的內積(inner product)「$\|q1\|\|p0\|cos\theta$」。當人臉 F1 與人臉 F0 之間的相似度越高時，角度「θ」越接近於零，特徵向量 p0 與特徵向量 q1 的內積接近於「$\|q1\|\|p0\|$」。因此，可根據特徵向量 p0 與特徵向量 q1 的乘積累加結果 MAC1 評估人臉 F1 與待辨識的人臉 F0 之間的相似度。

【0020】第 1B 圖為本揭示的記憶體裝置 1000 執行的 MAC 運算的流程圖。請參見第 1B 圖(並配合參見第 1A 圖的記憶體裝置 1000)，在步驟 S102，將待辨識的人臉 F0 的特徵向量 p0 輸入或饋入(feed)至記憶體裝置 1000。特徵向量 p0 可暫存於資料鎖存器 230。

【0021】而後，在步驟 S104，將特徵向量 p0 從資料鎖存器 230(即，共同資料鎖存器(CDL))轉移至資料鎖存器 210(即，第一資料鎖存器)。而後，在步驟 S106，執行選擇位元線讀取操作，以從記憶體陣列 100 的字元線 WL1~WLN 中分別讀取出特徵向量 q1~qN，並對其執行乘積運算。選擇位元線讀取操作得到的輸出

特徵向量 q1'~qN'為特徵向量 q1~qN 與特徵向量 p0 的乘積。輸出特徵向量 q1'~qN'可依序暫存於資料鎖存器 220。

【0022】而後，在步驟 S108，分別對於資料鎖存器 220 依序暫存的輸出特徵向量 q1'~qN'執行計數操作以得到計數結果 C1。而後，在步驟 S110，計數結果 C1 可轉移至資料鎖存器 230，並可經由資料鎖存器 230 輸出計數結果 C1。

【0023】另一方面，請參見表 1 及表 2。表 1 所示為一比較例的常規的 2 位元 MLC 的資料配置(data allocation) DA_1。記憶體陣列的一個實體頁面可包括多個邏輯頁面，例如高頁面(high page)及低頁面(low page)。以常規的 2 位元 MLC 為例，高頁面對應至 MLC 記憶單元的第一位元 CB1，低頁面對應至 MLC 記憶單元的第二位元 CB2。在比較例的資料配置 DA_1 中，不同的人臉的特徵向量可分別儲存於高頁面或低頁面，例如，人臉 F1~FN 之中的第 1 張人臉 F1 的特徵向量 q1 儲存於高頁面，第 2 張人臉 F2 的特徵向量 q2 儲存於低頁面。

表 1 (資料配置 DA_1)

高頁面 (第一位元 CB1)	人臉 F1 的特徵向量 q1
低頁面 (第二位元 CB2)	人臉 F2 的特徵向量 q2

【0024】表 2 所示為本揭示的 1.5 位元 MLC 的資料配置

DA_2。可將人臉的特徵向量其中每 8 個位元的資料區分為高有效位元(most significant bit，MSB)與低有效位元(least significant bit，LSB)的向量(vector)。例如，將人臉 F1 的特徵向量 q1 區分為 MSB 向量與 LSB 向量，MSB 向量與 LSB 向量例如各具有 4 個位元。MSB 向量的資料重要性較高而儲存於高頁面，LSB 向量的資料重要性較低而儲存於低頁面。類似的，下一個人臉 F2 的特徵向量 q2 的 MSB 向量儲存於高頁面，特徵向量 q2 的 LSB 向量儲存於低頁面。並且，高頁面對應至本揭示的 1.5 位元 MLC 的第一位元 CB1，低頁面對應至 1.5 位元 MLC 的第二位元 CB2。

表 2 (資料配置 DA_2)

高頁面 (第一位元 CB1)	人臉 F1 的特徵向量 q1 的 MSB 向量	人臉 F2 的特徵向量 q2 的 MSB 向量
低頁面 (第二位元 CB2)	人臉 F1 的特徵向量 q1 的 LSB 向量	人臉 F2 的特徵向量 q2 的 LSB 向量

【0025】第 2A 圖為常規的 2 位元 MLC 的臨界電壓分布的狀態的示意圖。第 2B 圖為本揭示一實施例的 1.5 位元 MLC 的臨界電壓分布的狀態的示意圖，可說明 1.5 位元 MLC 的機制。

【0026】請先參見第 2A 圖，可將記憶單元的臨界電壓 Vt 編程為不同的電壓位準，根據編程後的臨界電壓 Vt 的分布，常規的

2 位元 MLC 可具有 4 個狀態 S1’~S4’。狀態 S1’~S4’的數量「4」等於 2 的 2 次方。

【0027】其中，狀態 S1’介於 0v(零電位)與讀取電壓 V_{R1} 之間，狀態 S2’介於讀取電壓 V_{R1} 與讀取電壓 V_{R2} 之間，狀態 S3’介於讀取電壓 V_{R2} 與讀取電壓 V_{R3} 之間，狀態 S4’介於讀取電壓 V_{R3} 與通過電壓 V_{Pass} 之間。常規的 2 位元 MLC 可儲存 2 個位元的資料(CB1,CB2)。其中，狀態 S1’對應於資料(CB1=1,CB2=1)，狀態 S2’對應於資料(CB1=1,CB2=0)，狀態 S3’對應於資料(CB1=0,CB2=0)，狀態 S4’對應於資料(CB1=0,CB2=1)。常規的 2 位元 MLC 的狀態 S1’~S4’對應的電壓間距 W1’~W4’較窄，可能導致資料(CB1,CB2)的讀取錯誤，進而減損記憶體內運算(例如 MAC 運算)的運算精確度。

【0028】請參見第 2B 圖，以本揭示的 1.5 位元 MLC 而言，臨界電壓 Vt 的分布僅區分為 N 個狀態，N 小於 2 的 2 次方。例如，N=3，則 1.5 位元 MLC 具有 3 個狀態 S1~S3。其中，狀態 S1 介於 0v(零電位)與讀取電壓 V_{R1} 之間，狀態 S2 介於讀取電壓 V_{R1} 與讀取電壓 V_{R2} 之間，狀態 S3 介於讀取電壓 V_{R2} 與通過電壓 V_{Pass} 之間。1.5 位元 MLC 可儲存的資料之等效位元數為 1.5 位元，其中，狀態 S1 對應於資料(CB1=1,CB2=1)，狀態 S2 對應於資料(CB1=1,CB2=0)，狀態 S3 對應於資料(CB1=0,CB2=0)。換言之，1.5 位元 MLC 不包括資料(CB1=0,CB2=1)，可將常規的 2 位元 MLC 的狀態 S3’與狀態 S4’視為 1.5 位元 MLC 的單一個狀態 S3。

1.5 位元 MLC 的狀態 S1~S3 對應的電壓間距 W1~W3 大於常規的 2 位元 MLC 的電壓間距 W1'~W4'，1.5 位元 MLC 的資料(CB1,CB2)的讀取正確率可因此提升。

【0029】表 3 所示為第一位元 CB1 與第二位元 CB2 的內容的一種示例，可顯示 1.5 位元 MLC 的資料(CB1,CB2)與常規的 2 位元 MLC 之區別。

表 3

常規的 2 位元 MLC	
高頁面	第一位元 CB1 10010100
低頁面	第二位元 CB2 01011011
1.5 位元 MLC	
高頁面	第一位元 CB1 10010100
低頁面	第二位元 CB2 00010000

【0030】本揭示的記憶體裝置 1000 除了採用 1.5 位元之 MLC 以提升資料的讀取正確率，在運作上，記憶體裝置 1000 可減少冗餘(redundant)的位元線設定(BL setup)的操作，以減少資料的讀取操作所需的時間。

【0031】對於常規的 2 位元 MLC 而言，低頁面的第二位元

CB2 的讀取操作需執行一次的位元線設定並施加一次的讀取電壓 V_{R2}。一次的位元線設定時間 t_BL_S 例如為 60μs，施加一次的讀取電壓 V_{R2} 對應的讀取時間 t_Vread 例如為 15μs。即，第二位元 CB2 的讀取操作的執行時間大致為 60μs+15μs=75μs。另一方面，高頁面的第一位元 CB1 的讀取操作需執行一次的位元線設定並施加兩次的讀取電壓 V_{R1}、V_{R3}。即，第一位元 CB1 的讀取操作的執行時間大致為 60μs+2*15μs=90μs。由上，常規的 2 位元 MLC 的高頁面與低頁面的資料(CB1,CB2)的讀取時間大致為 75μs+90μs=165μs。

【0032】此外，對於常規的 1 位元的單階記憶單元(single-level cell，SLC)而言，由於常規的 1 位元 SLC 僅儲存 1 個位元的資料，讀取 2 個位元的資料需要執行兩次讀取操作。一次讀取操作需執行一次的位元線設定並施加一次的讀取電壓 V_{R1}，執行時間大致為 60μs+15μs=75μs。因此，常規的 1 位元 SLC 讀取 2 個位元的資料的讀取時間大致為 2*75μs=150μs。

【0033】相對地，本揭示的記憶體裝置 1000 在進行第一位元 CB1 及第二位元 CB2 的讀取操作時僅執行一次的位元線設定，因此僅需耗費一次的位元線設定時間 t_BL_S=60μs。並且，本揭示的記憶體裝置 1000 採用的 1.5 位元 MLC 僅操作於 2 個讀取電壓 V_{R1}、V_{R2}，因此僅需耗費兩次的讀取時間 t_Vread，即 2*15μs=30μs。由上，1.5 位元 MLC 的高頁面與低頁面的資料(CB1,CB2)的讀取時間大致為 60μs+30μs=90μs。據此，相較於

常規的 2 位元 MLC 的讀取時間 165μs 及常規的 1 位元 SLC 的讀取時間 150μs，本揭示的記憶體裝置 1000 執行僅執行一次的位元線設定並配合於 1.5 位元 MLC 的機制而在讀取時間延遲(latency)上能夠大幅改善。

【0034】第 3 圖為本揭示的記憶體裝置 1000 執行的 MAC 運算之另一實施例的流程圖。請參見第 3 圖，在步驟 S202 中，判斷記憶體裝置 1000 的狀況(condition)。當判斷記憶體裝置 1000 將執行 MAC 運算時，接著進行步驟 S204：從 1.5 位元 MLC 組成的記憶體陣列 100 中讀取特徵向量 qi 的資料。例如，執行選擇位元線讀取操作的指令，執行一次位元線設定並施加兩次讀取電壓 V_{R1}、V_{R2} 以讀取 1.5 位元 MLC 的每一個資料(CB1,CB2)。記憶體陣列 100 的高頁面儲存的第一位元 CB1 對應於特徵向量 qi 的 MSB 向量，低頁面儲存的第二位元 CB2 對應於特徵向量 qi 的 LSB 向量。可對於 MSB 向量與 LSB 向量分別執行不同資料量的計數操作，以進行累加運算。

【0035】另一方面，在步驟 S202 中，當判斷記憶體裝置 1000 將移除(remove)資料時，接著進行步驟 S210：刪除資料。例如，刪除記憶體陣列 100 儲存的 N 個特徵向量 q1~qN。

【0036】另一方面，在步驟 S202 中，當判斷記憶體裝置 1000 將產生(create)資料、插入(insert)資料或更新(update)資料時，接著進行步驟 S206：量化(quantize)並計算修正後(revised)的特徵向量 qi。接著，進行步驟 S208：將修正後的特徵向量 qi 的資

料編程寫入(program)記憶體陣列 100。特徵向量 qi 的資料的寫入操作可根據第 2B 圖所示的 1.5 位元 MLC 的資料樣式(data pattern)及表 2 所示的資料配置而執行。

【0037】第4A~4D圖為本揭示的不同實施例的1.5位元MLC的臨界電壓分布的狀態的示意圖，分別說明1.5位元MLC的不同資料樣式。請先參見第4A圖，在1.5位元MLC的第一資料樣式DP_1中，臨界電壓Vt之分布的狀態S1對應於資料(CB1=1,CB2=1)，狀態S2對應於資料(CB1=1,CB2=0)，狀態S3對應於資料(CB1=0,CB2=0)。常規的2位元MLC的狀態S3'及狀態S4'可視為1.5位元MLC的單一的狀態S3，並且，1.5位元MLC的第一資料樣式DP_1不包括資料(CB1=0,CB2=1)，而是將資料(CB1=0,CB2=1)其中的第二位元CB2=「1」視為「0」。由於第二位元CB2對應於低頁面以儲存資料重要性較低的LSB向量，將第二位元CB2=「1」視為「0」不會影響資料精確度。

【0038】請參見第4B圖，在1.5位元MLC的第二資料樣式DP_2中，臨界電壓Vt之分布的狀態S1對應於資料(CB1=1,CB2=1)，狀態S2對應於資料(CB1=1,CB2=0)，狀態S3對應於資料(CB1=0,CB2=1)。常規的2位元MLC的狀態S3'及狀態S4'可視為1.5位元MLC的單一個狀態S3，並且，1.5位元MLC的第二資料樣式DP_2不包括資料(CB1=0,CB2=0)，而是將資料(CB1=0,CB2=0)其中的第二位元CB2=0視為第二位元CB2=1。

【0039】請參見第 4C 圖，在 1.5 位元 MLC 的第三資料樣式 DP_3 中，臨界電壓 Vt 之分布的狀態 S1 對應於資料(CB1=1,CB2=0)，狀態 S2 對應於資料(CB1=0,CB2=0)，狀態 S3 對應於資料(CB1=0,CB2=1)。常規的 2 位元 MLC 的狀態 S1'及狀態 S2'可視為 1.5 位元 MLC 的單一的狀態 S1，並且，1.5 位元 MLC 的第三資料樣式 DP_3 不包括資料(CB1=1,CB2=1)，而是將資料(CB1=1,CB2=1)其中的第二位元 CB2=1 視為第二位元 CB2=0。

【0040】請參見第 4D 圖，在 1.5 位元 MLC 的第四資料樣式 DP_4 中，臨界電壓 Vt 之分布的狀態 S1 對應於資料(CB1=1,CB2=1)，狀態 S2 對應於資料(CB1=0,CB2=0)，狀態 S3 對應於資料(CB1=0,CB2=1)。常規的 2 位元 MLC 的狀態 S1'及狀態 S2'可視為 1.5 位元 MLC 的單一的狀態 S1，並且，1.5 位元 MLC 的第四資料樣式 DP_4 不包括資料(CB1=1,CB2=0)，而是將資料(CB1=1,CB2=0)其中的第二位元 CB2=「0」視為「1」。

【0041】第 4A~4D 圖所示的 1.5 位元 MLC 的第一資料樣式 DP_1 至第四資料樣式 DP_4 可配合於表 2 所示的資料配置 DA_2 而實施於本揭示的記憶體裝置 1000，用以儲存特徵向量 qi 的 MSB 向量及 LSB 向量以執行 MAC 運算。在第一資料樣式 DP_1 至第四資料樣式 DP_4 中，第二位元 CB2=「0」視為「1」。由於第二位元 CB2 對應於低頁面以儲存資料重要性較低的 LSB 向量，第二位元 CB2=「0」視為「1」的影響不大。

【0042】在其他示例中，本揭示的記憶體裝置 1000 的記憶單元亦可為三階記憶單元(triple level cell，TLC)，TLC 記憶單元可具有第一位元 CB1、第二位元 CB2 及第三位元 CB3。請先參見第 5A 圖的常規的 3 位元 TLC 記憶單元的臨界電壓分布的狀態的示意圖。常規的 3 位元 TLC 可具有 8 個狀態 S1’~S8’而對應至 3 個位元的資料(CB1,CB2,CB3)。狀態 S1’~S8’的數量「8」等於 2 的 3 次方。其中，狀態 S1’對應於資料(CB1=1,CB2=1,CB3=1)，狀態 S2’對應於資料(CB1=1,CB2=1,CB3=0)，狀態 S3’對應於資料(CB1=1,CB2=0,CB3=0)，狀態 S4’對應於資料(CB1=0,CB2=0,CB3=0)，狀態 S5’對應於資料(CB1=0,CB2=1,CB3=0)，狀態 S6’對應於資料(CB1=0,CB2=1,CB3=1)，狀態 S7’對應於資料(CB1=0,CB2=0,CB3=1)，狀態 S8’對應於資料(CB1=1,CB2=0,CB3=1)。狀態 S1’~S8’分別對應至較窄的電壓間距 W1’~W8’，可能導致資料(CB1,CB2,CB3)的讀取錯誤。

【0043】請參見第 5B 圖繪示之本揭示一實施例的 TLC 記憶單元的臨界電壓分布的狀態的示意圖，本實施例的 TLC 的臨界電壓 Vt 分布僅區分為 M 個狀態，M 小於 2 的 3 次方。例如，M=5，則本實施例的 TLC 具有 5 個狀態 S1~S5。。其中，狀態 S1 對應於資料(CB1=1,CB2=1,CB3=1)，狀態 S2 對應於資料(CB1=1,CB2=0,CB3=0)，狀態 S3 對應於資料(CB1=0,CB2=0,CB3=0)，狀態 S4 對應於資料

(CB1=0,CB2=1,CB3=1)，狀態 S5 對應於資料(CB1=0,CB2=0,CB3=1)。本實施例的 TLC 儲存的資料不包括資料(CB1=1,CB2=1,CB3=0)，可將常規的 3 位元 TLC 的狀態 S1’與狀態 S2’視為本實施例的 TLC 的單一個狀態 S1。類似的，將常規的 3 位元 TLC 的狀態 S4’與狀態 S5’視為本實施例的 TLC 的單一個狀態 S3，並排除資料(CB1=0,CB2=1,CB3=0)。將常規的 3 位元 TLC 的狀態 S7’與狀態 S8’視為本實施例的 TLC 的單一個狀態 S5 而排除資料(CB1=1,CB2=0,CB3=1)。本實施例的 TLC 儲存的資料之等效位元數雖小於 3 位元，然而本實施例的 TLC 的狀態 S1~S5 對應的電壓間距較大，可提升資料(CB1,CB2,CB3)的讀取正確率。第 5B 圖所示的狀態 S1~S5 與其對應的資料(CB1,CB2,CB3)的樣式可稱為 TLC 的第一資料樣式 DP_1’。

【0044】第 5C~5I 圖為本揭示的不同實施例的 TLC 的臨界電壓分布的狀態的示意圖，分別說明 TLC 的不同資料樣式。

【0045】首先，第 5C 圖所示的第二資料樣式 DP_2’類似於第一資料樣式 DP_1’，差異處在於：將常規的 3 位元 TLC 的狀態 S7’與狀態 S8’視為本實施例的 TLC 的單一個狀態 S5 而排除資料(CB1=0,CB2=0,CB3=1)。

【0046】根據不同的資料排列，可得到 TLC 的其他資料樣式。第 5D 圖所示的第三資料樣式 DP_3’類似於第一資料樣式 DP_1’，差異處在於：排除資料(CB1=0,CB2=0,CB3=0)。第 5E 圖所示的

第四資料樣式 DP_4’類似於第一資料樣式 DP_1’，差異處在於：排除資料(CB1=0,CB2=1,CB3=1)。第 5F 圖所示的第五資料樣式 DP_5’類似於第三資料樣式 DP_3’，差異處在於：排除資料(CB1=0,CB2=0,CB3=1)。第 5G 圖所示的第六資料樣式 DP_6’類似於第四資料樣式 DP_4’，差異處在於：排除資料(CB1=0,CB2=0,CB3=1)。第 5H 圖所示的第七資料樣式 DP_7’類似於第四資料樣式 DP_4’，差異處在於：排除資料(CB1=0,CB2=0,CB3=0)。第 5I 圖所示的第八資料樣式 DP_8’類似於第七資料樣式 DP_7’，差異處在於：排除資料(CB1=0,CB2=0,CB3=1)。

【0047】另一方面，請參見表 4 及表 5。表 4 所示為另一比較例的常規的 3 位元 TLC 的資料配置 DA_3。記憶體陣列的一個實體頁面可包括高頁面、中頁面及低頁面，分別對應至 TLC 記憶單元的第一位元 CB1、第二位元 CB2 及第三位元 CB3。人臉 F1~FN 之中的第 1 張人臉 F1 的特徵向量 q1 儲存於高頁面，第 2 張人臉 F2 的特徵向量 q2 儲存於中頁面，第 3 張人臉 F3 的特徵向量 q3 儲存於低頁面。

表 4 (資料配置 DA_3)

高頁面 (第一位元 CB1)	人臉 F1 的特徵向量 q1
中頁面	人臉 F2 的特徵向量 q2

(第二位元 CB2)	
低頁面 (第三位元 CB3)	人臉 F3 的特徵向量 q3

【0048】表 5 所示為本揭示的 TLC 的資料配置 DA_4。可將每一個特徵向量的資料區分為第一部分 P1、第二部分 P2 及第三部分 P3，根據各部分的資料重要性程度而分別儲存於高頁面、中頁面及低頁面。

表 5 (資料配置 DA_4)

高頁面 (第一位元 CB1)	人臉 F1 的特徵向量 q1 的第一部分 P1	人臉 F2 的特徵向量 q2 的第一部分 P1	人臉 F3 的特徵向量 q3 的第一部分 P1
中頁面 (第二位元 CB2)	人臉 F1 的特徵向量 q1 的第二部分 P2	人臉 F2 的特徵向量 q2 的第二部分 P2	人臉 F3 的特徵向量 q3 的第二部分 P2
低頁面 (第三位元 CB3)	人臉 F1 的特徵向量 q1 的第三部分 P3	人臉 F2 的特徵向量 q2 的第三部分 P3	人臉 F3 的特徵向量 q3 的第三部分 P3

【0049】上述實施例說明等效位元數較少的 MLC 與 TLC 的各種資料樣式及資料配置。類似的機制亦可應用於四階記憶單元 (quad level cell，QLC)。

【0050】雖然本發明已以較佳實施例及範例詳細揭露如上，可理解的是，此些範例意指說明而非限制之意義。可預期的是，

所屬技術領域中具有通常知識者可想到多種修改及組合，其多種修改及組合落在本發明之精神以及後附之申請專利範圍之範圍內。

【符號說明】

【0051】

1000:記憶體裝置

100:記憶體陣列

210:第一資料鎖存器

220:第二資料鎖存器

230:共同資料鎖存器

WL1~WLN:字元線

p0,q1',q1~qN:特徵向量

S1~S5,S1'~S8':狀態

V_{R1}~V_{R7}:讀取電壓

V_{Pass}:通過電壓

W1~W3:電壓間隔

W1'~W4':電壓間隔

DP_1:第一資料樣式

DP_2:第二資料樣式

DP_3:第三資料樣式

DP_4:第四資料樣式

DP_1':第一資料樣式

DP_2':第二資料樣式

DP_3’:第三資料樣式

DP_4’:第四資料樣式

DP_5’:第五資料樣式

DP_6’:第六資料樣式

DP_7’:第七資料樣式

DP_8’:第八資料樣式

Vt:臨界電壓

S102~S110:步驟

S202~S210:步驟

【發明申請專利範圍】

【請求項 1】一種記憶體裝置，包括：

一記憶體陣列，用於儲存複數個向量資料，各該向量資料具有一高有效位元(MSB)向量與一低有效位元(LSB)向量，該記憶體陣列包括：

複數個記憶單元，各該記憶單元具有一第一位元及一第二位元，該第一位元用於儲存各該向量資料的該 MSB 向量，該第二位元用於儲存各該向量資料的該 LSB 向量；

其中，各該向量資料對應的一位元線執行一次的位元線設定，並根據該位元線讀取各該向量資料的該 MSB 向量與該 LSB 向量，並且各該記憶單元的臨界電壓分布區分為 N 個狀態，N 為正整數且 N 小於 2 的 2 次方，各該記憶單元儲存的等效位元數小於 2。

【請求項 2】如請求項 1 所述之記憶體裝置，其中該記憶體陣列的一實體頁面具有複數個邏輯頁面，該些邏輯頁面包括一高頁面及一低頁面，該高頁面對應於該第一位元以儲存該 MSB 向量，該低頁面對應於該第二位元以儲存該 LSB 向量。

【請求項 3】如請求項 1 所述之記憶體裝置，其中各該記憶單元為多階記憶單元，各該記憶單元的臨界電壓分布區分為一第一狀態、一第二狀態及一第三狀態；

其中，該第一狀態僅對應於第一位元為「1」且第二位元為「0」

的資料，或該第一狀態僅對應於第一位元為「1」且第二位元為「1」的資料。

【請求項 4】如請求項 1 所述之記憶體裝置，其中各該記憶單元為二階記憶單元，各該記憶單元的臨界電壓分布區分為一第一狀態、一第二狀態及一第三狀態；

其中，該第三狀態僅對應於第一位元為「0」且第二位元為「0」的資料，或該第三狀態僅對應於第一位元為「0」且第二位元為「1」的資料。

【請求項 5】如請求項 1 所述之記憶體裝置，其中各該記憶單元為三階記憶單元，各該記憶單元更具有一第三位元，該第一位元用於儲存各該向量資料的一第一部分，該第二位元用於儲存各該向量資料的一第二部分，且該第三位元用於儲存各該向量資料的一第三部分；

其中，各該記憶單元的臨界電壓分布區分為 N 個狀態，N 為正整數且 N 小於 2 的 3 次方，各該記憶單元儲存的等效位元數小於 3。

【請求項 6】如請求項 1 所述之記憶體裝置，其中該些向量資料包括複數個權重特徵向量及一輸入特徵向量，並且該記憶體陣列包括：

複數條字元線，用於儲存該些權重特徵向量；

其中，該記憶體裝置執行一選擇位元線讀取操作，以從該些字元線分別讀取該些權重特徵向量。

【請求項 7】如請求項 6 所述之記憶體裝置，更包括：

一共同資料鎖存器，用於接收該輸入特徵向量；以及

一第一資料鎖存器，用於暫存該輸入特徵向量；

其中，該記憶體裝置根據該輸入特徵向量與該些權重特徵向量的其中一者執行一乘積運算。

【請求項 8】如請求項 7 所述之記憶體裝置，更包括：

一第二資料鎖存器，用於暫存該乘積運算的結果；

其中，該記憶體裝置根據該乘積運算的結果執行一累加運算。

【請求項 9】一種記憶體裝置的操作方法，其中該記憶體裝置包括一記憶體陣列，該記憶體陣列包括複數個記憶單元，各該記憶單元具有一第一位元及一第二位元，該操作方法包括：

儲存複數個向量資料於該記憶體陣列，各該向量資料具有一高有效位元(MSB)向量與一低有效位元(LSB)向量；

儲存各該向量資料的該 MSB 向量於該第一位元；

儲存各該向量資料的該 LSB 向量於該第二位元；

對於各該向量資料對應的一位元線執行一次的位元線設定；以

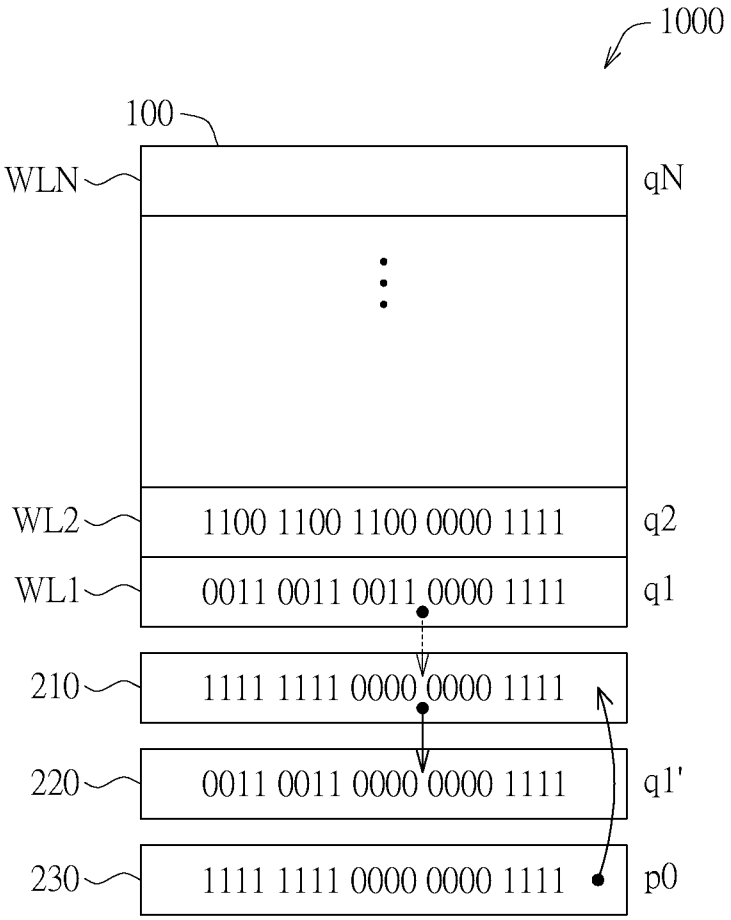
及

根據該位元線讀取各該向量資料的該 MSB 向量與該 LSB 向量；

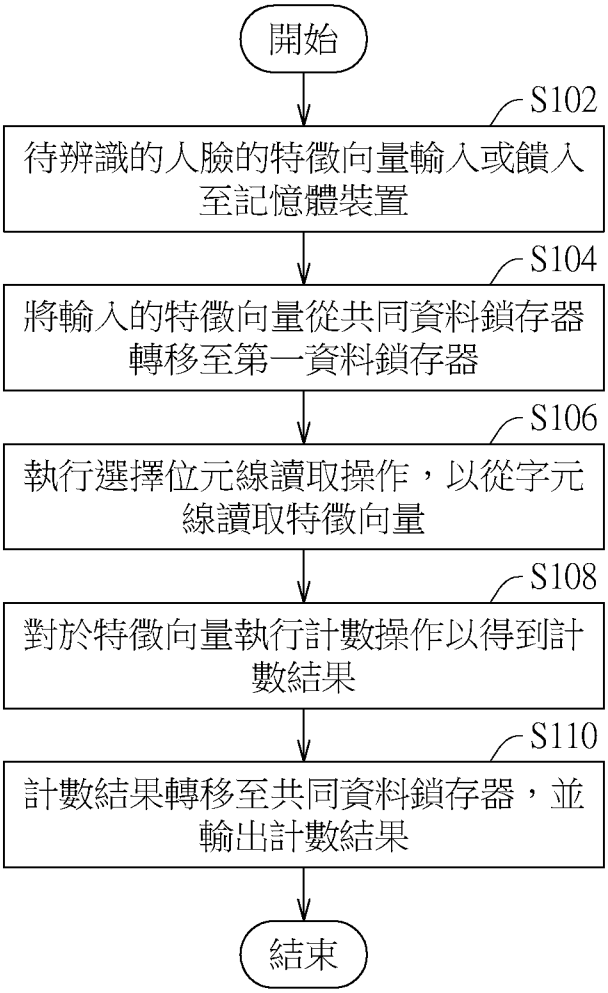
其中，各該記憶單元的臨界電壓分布區分為 N 個狀態，N 為正整數且 N 小於 2 的 2 次方，各該記憶單元儲存的等效位元數小於 2。

【請求項 10】如請求項 9 所述之操作方法，其中該記憶體陣列的一實體頁面具有複數個邏輯頁面，該些邏輯頁面包括一高頁面及一低頁面，該高頁面對應於該第一位元以儲存該 MSB 向量，該低頁面對應於該第二位元以儲存該 LSB 向量。

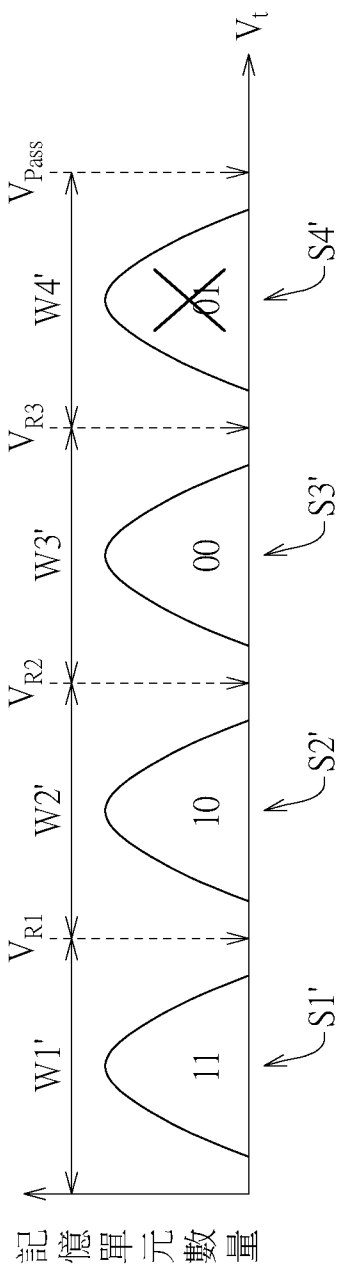
【發明圖式】


1000
100
WLN
qN
WL2
1100 1100 1100 0000 1111
q2
WL1
0011 0011 0011 0000 1111
q1
210
1111 1111 0000 0000 1111
220
0011 0011 0000 0000 1111
q1'
230
1111 1111 0000 0000 1111
p0

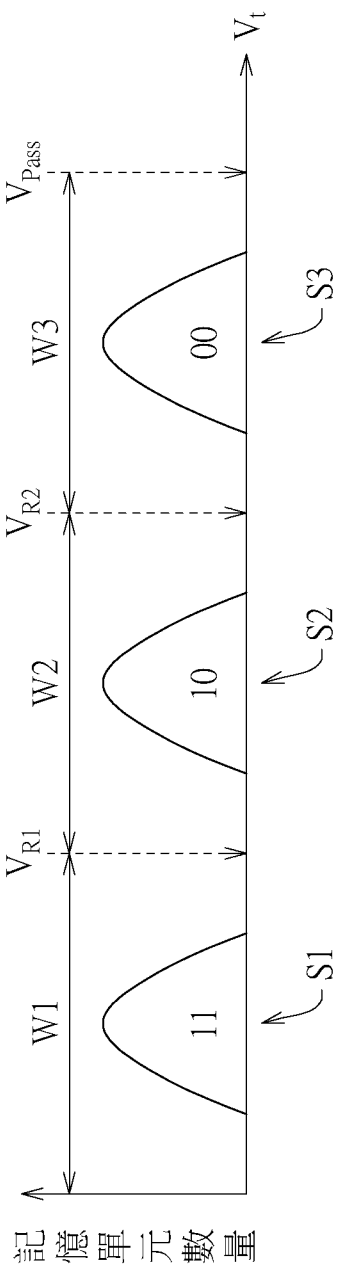
第 1A 圖


開始
S102
待辨識的人臉的特徵向量輸入或饋入至記憶體裝置
S104
將輸入的特徵向量從共同資料鎖存器轉移至第一資料鎖存器
S106
執行選擇位元線讀取操作，以從字元線讀取特徵向量
S108
對於特徵向量執行計數操作以得到計數結果
S110
計數結果轉移至共同資料鎖存器，並輸出計數結果
結束

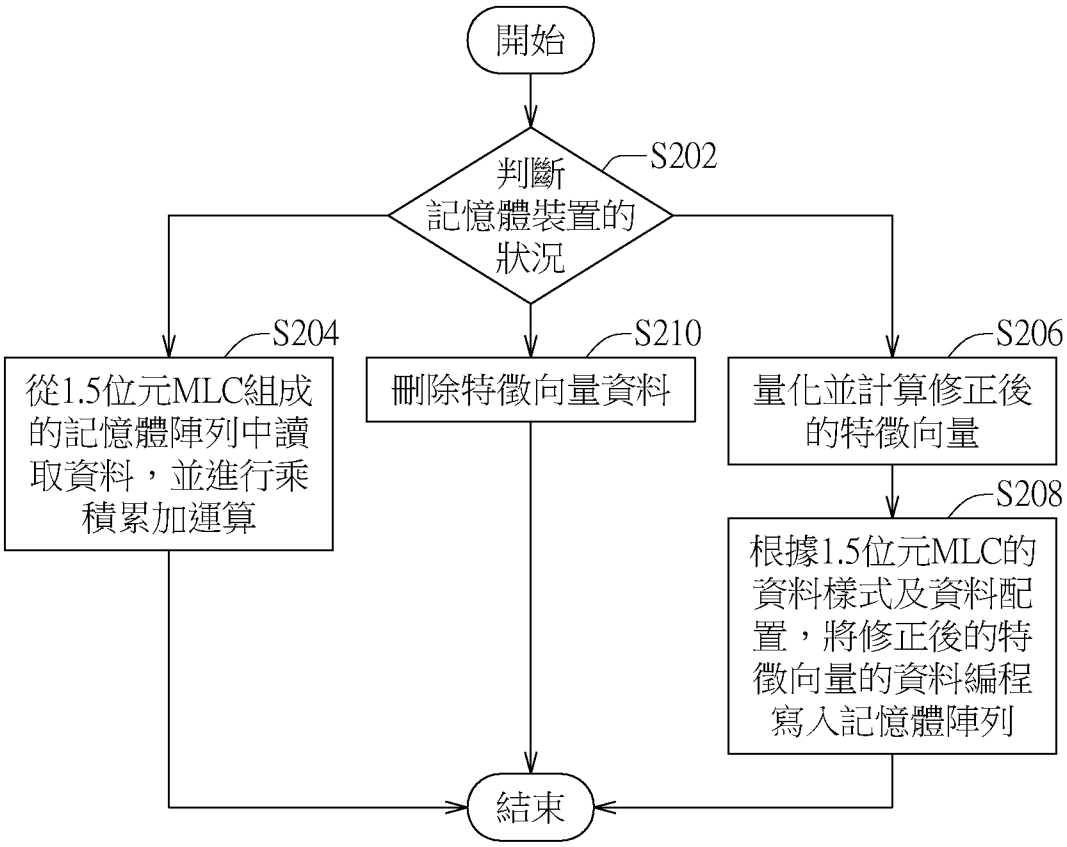
第 1B 圖


記憶單元數量
V_t
V_{Pass}
V_{R1}
V_{R2}
V_{R3}
W1'
W2'
W3'
W4'
11
10
00
01
S1'
S2'
S3'
S4'

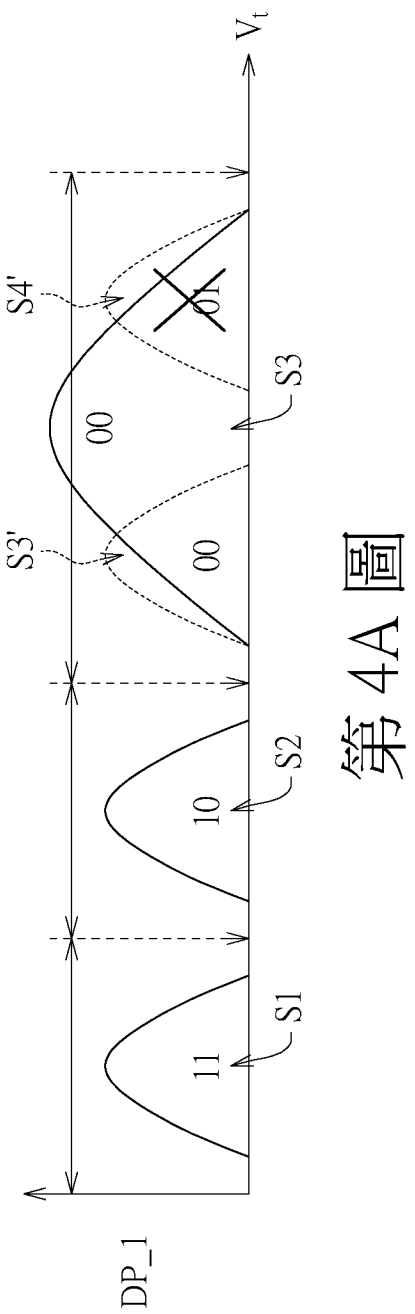
第 2A 圖


記憶單元數量
V_t
V_{Pass}
V_{R1}
V_{R2}
W1
W2
W3
11
10
00
S1
S2
S3

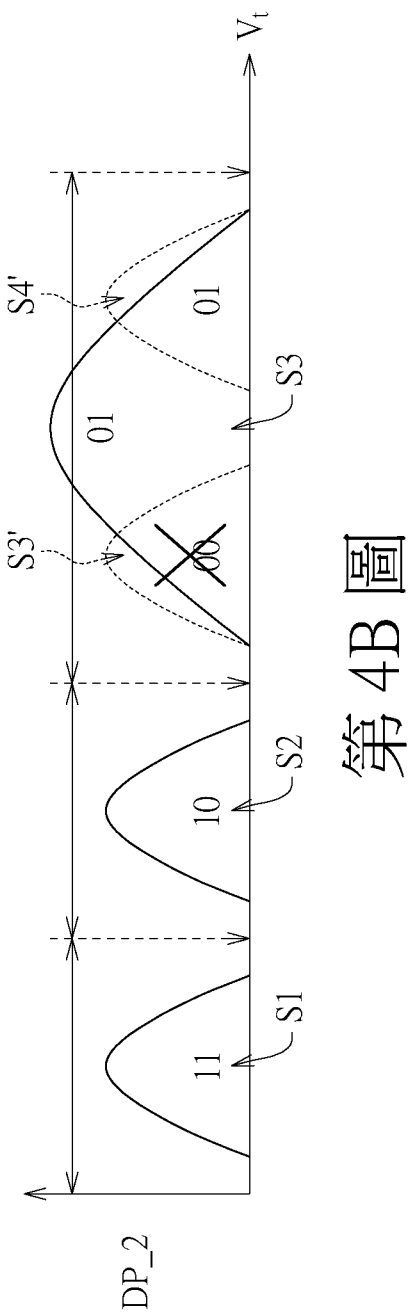
第 2B 圖


開始
判斷記憶體裝置的狀況
S202
S204
從1.5位元MLC組成的記憶體陣列中讀取資料，並進行乘積累加運算
S210
刪除特徵向量資料
S206
量化並計算修正後的特徵向量
S208
根據1.5位元MLC的資料樣式及資料配置，將修正後的特徵向量的資料編程寫入記憶體陣列
結束

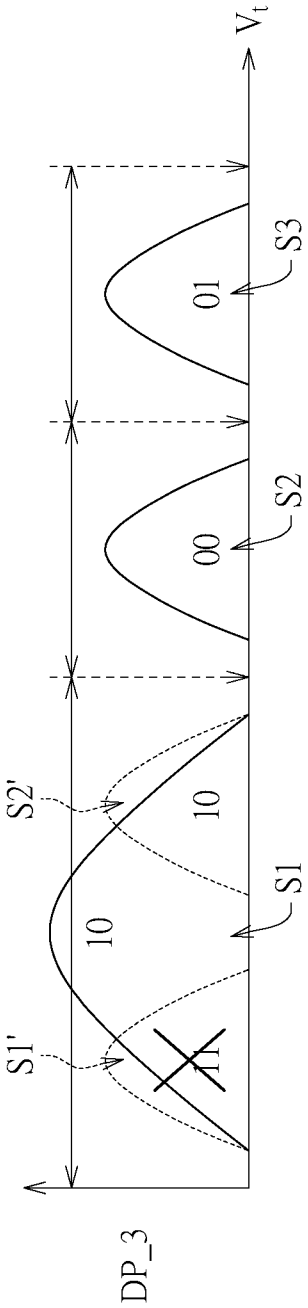
第 3 圖


V_t
S4'
S3'
S3
S2
S1
01
00
00
10
11
DP_1

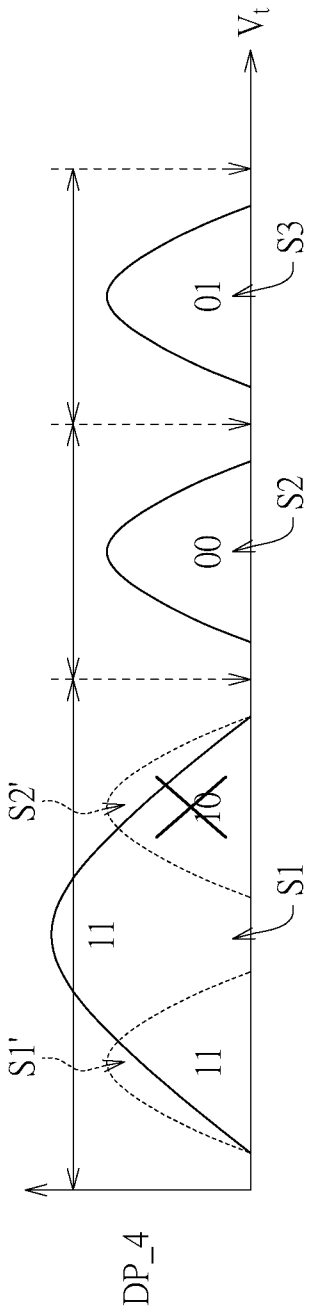
第 4A 圖


V_t
S4'
S3'
S3
S2
S1
01
01
00
10
11
DP_2

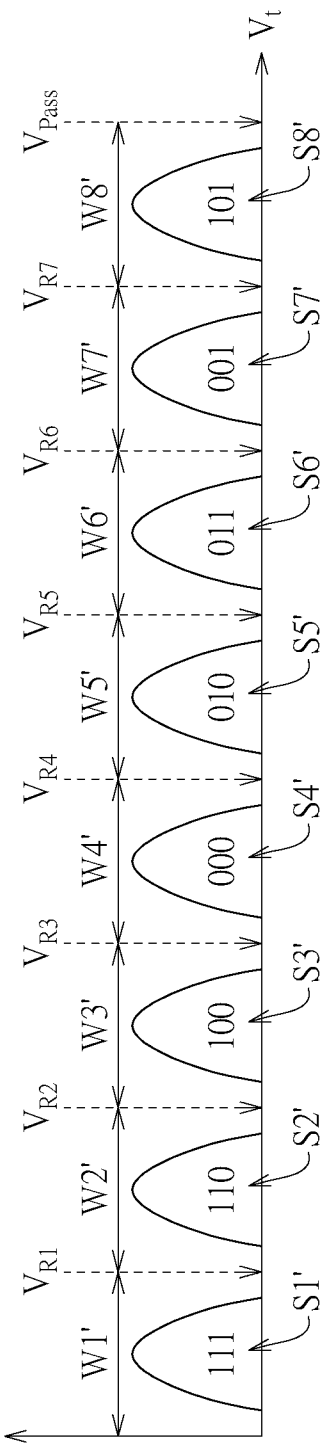
第 4B 圖


V_t
S3
01
S2
00
S2'
10
S1
10
S1'
11
DP_3

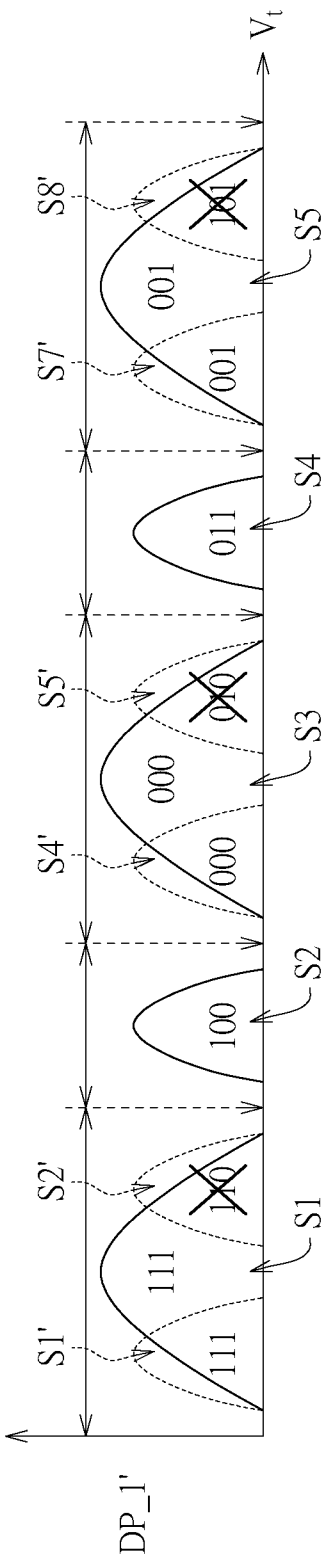
第 4C 圖


V_t
S3
01
S2
00
S2'
10
S1
11
S1'
11
DP_4

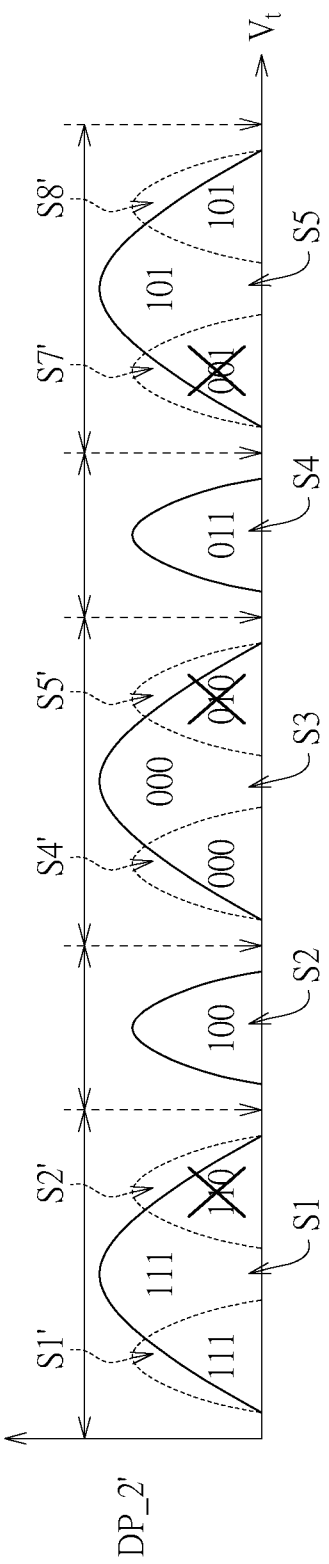
第 4D 圖


V_t
V_{Pass}
V_{R7}
V_{R6}
V_{R5}
V_{R4}
V_{R3}
V_{R2}
V_{R1}
W8'
W7'
W6'
W5'
W4'
W3'
W2'
W1'
S8'
S7'
S6'
S5'
S4'
S3'
S2'
S1'
101
001
011
010
000
100
110
111

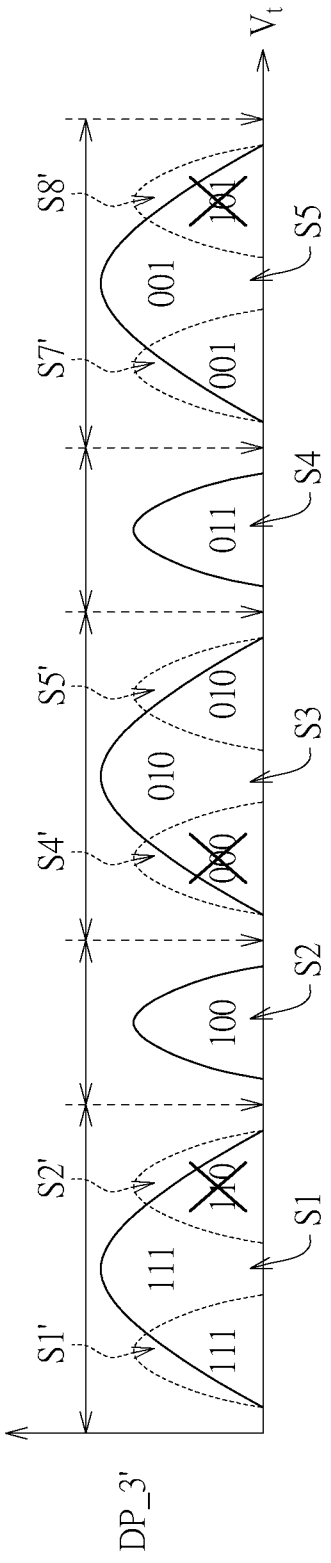
第 5A 圖


V_t
S8'
S7'
S5'
S4'
S2'
S1'
S5
S4
S3
S2
S1
101
001
011
000
100
111
DP_1'

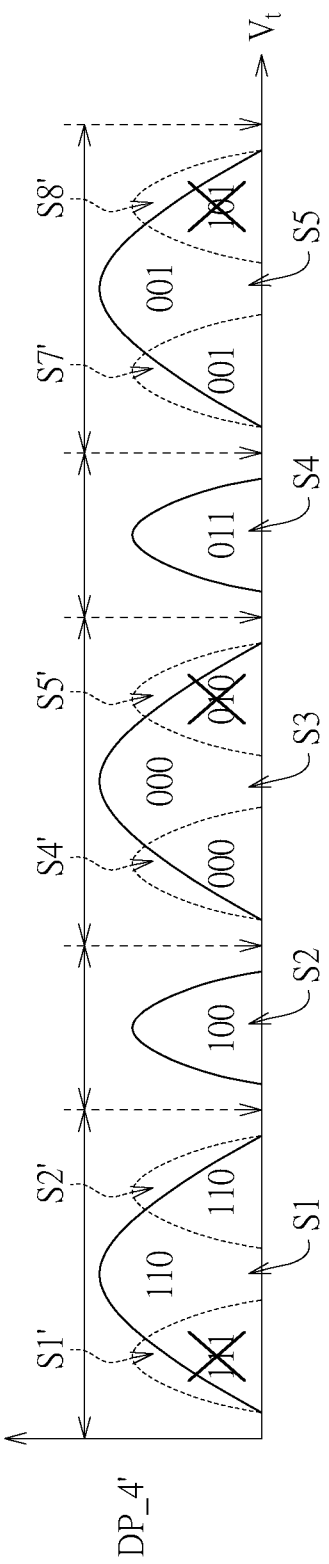
第 5B 圖


V_t
S1'
S2'
S4'
S5'
S7'
S8'
S1
S2
S3
S4
S5
111
110
100
000
010
011
001
101
DP_2'

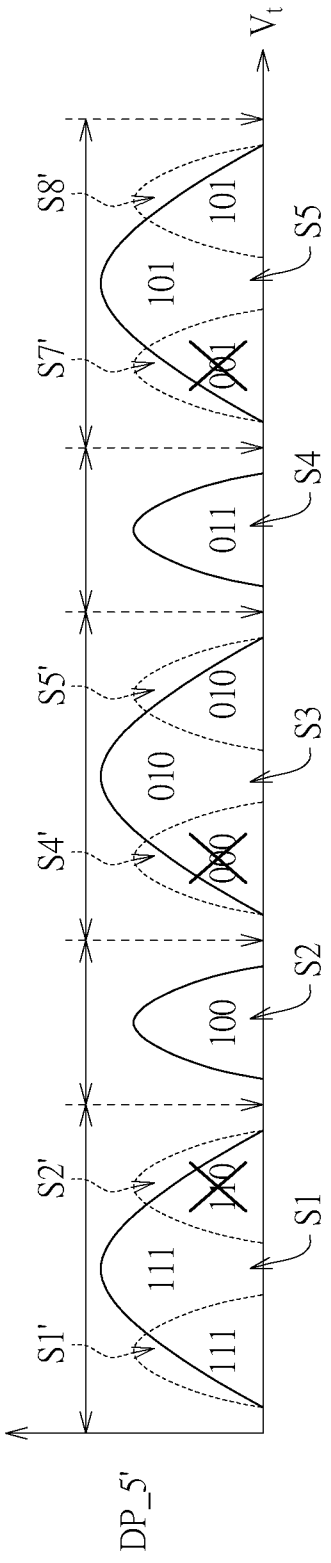
第 5C 圖


V_t
S1'
S2'
S4'
S5'
S7'
S8'
S1
S2
S3
S4
S5
111
110
100
000
010
011
001
101
DP_3'

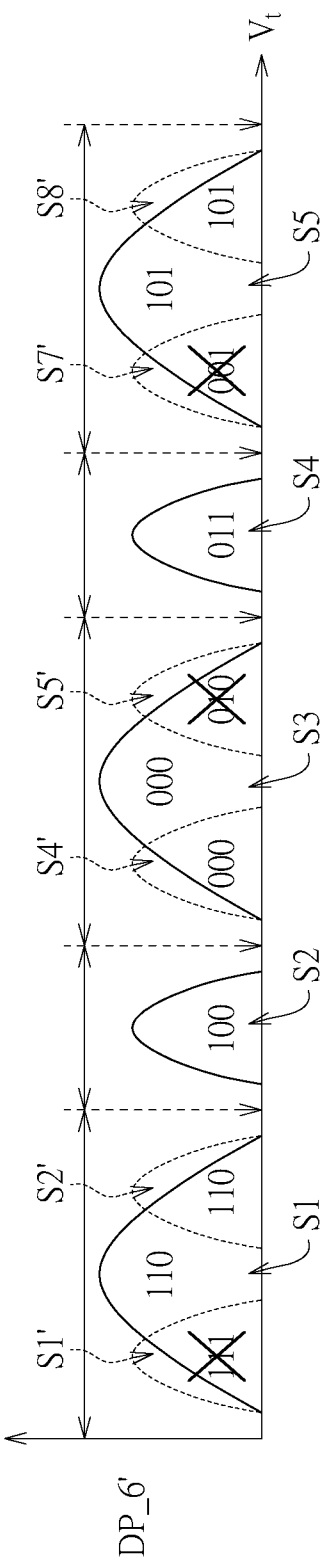
第 5D 圖


DP_4'
V_t
S1'
S2'
S4'
S5'
S7'
S8'
S1
S2
S3
S4
S5
111
110
100
000
010
011
001
101

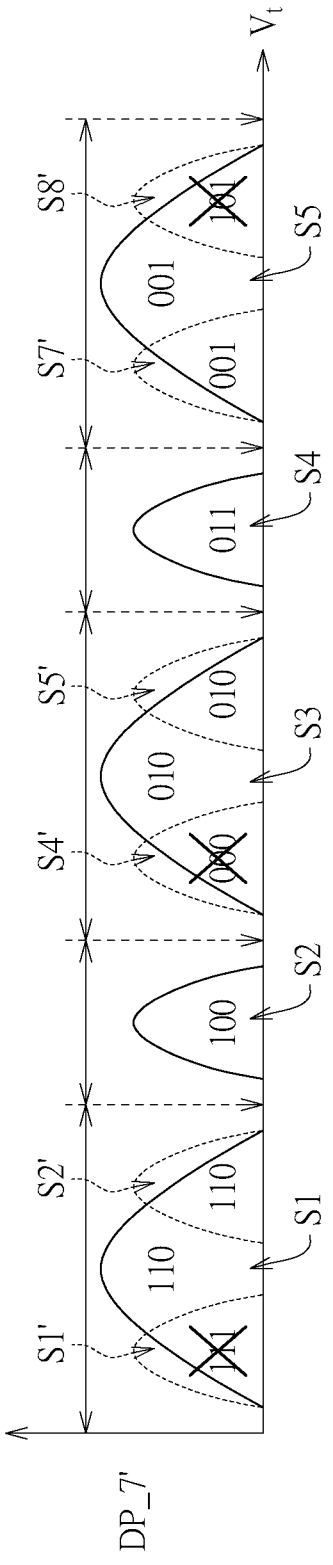
第 5E 圖


DP_5'
V_t
S1'
S2'
S4'
S5'
S7'
S8'
S1
S2
S3
S4
S5
111
110
100
000
010
011
001
101

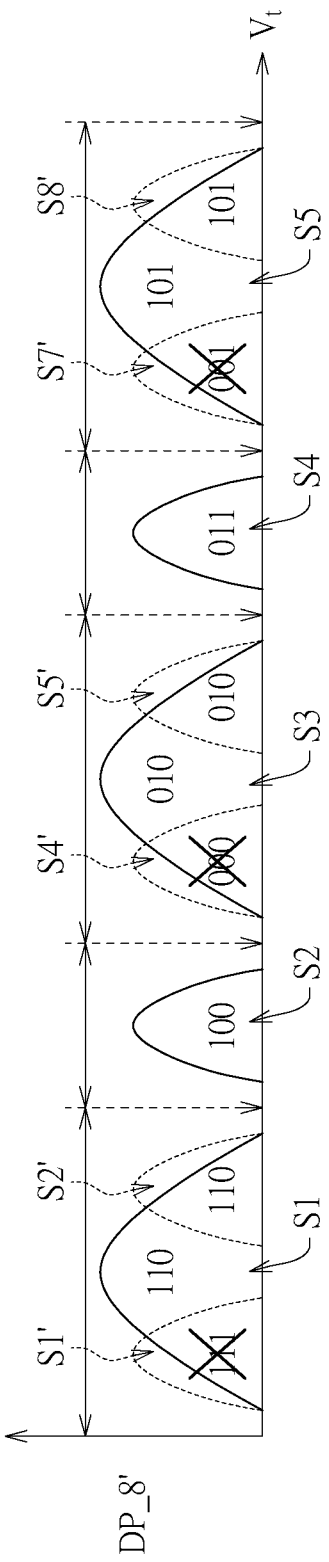
第 5F 圖


DP_6'
V_t
S1'
S2'
S4'
S5'
S7'
S8'
S1
S2
S3
S4
S5
111
110
110
100
000
000
010
011
001
101
101

第 5G 圖


DP_7'
V_t
S1'
S2'
S4'
S5'
S7'
S8'
S1
S2
S3
S4
S5
111
110
110
100
000
010
010
011
001
001
101

第 5H 圖


V_t
DP_8'
S1'
S2'
S4'
S5'
S7'
S8'
S1
S2
S3
S4
S5
101
101
001
011
010
010
000
100
110
110
111

第 5I 圖