國立臺灣大學電機資訊學院資訊網路與多媒體研究所

碩士論文

Graduate Institute of Networking and Multimedia
College of Electrical Engineering and Computer Science
National Taiwan University

Master Thesis

透過穿戴式攝影機評估在社群中的視覺人類簽章認證

An Evaluation of Visual Human Signature Identification in Community via Wearable Camera

曹嘉慶
Chia-Chin Tsao

指導教授：徐宏民博士
Advisor: Winston H. Hsu, Ph.D.

中華民國 103 年 7 月
July, 2014

國立臺灣大學碩士學位論文
口試委員會審定書

透過穿戴式攝影機評估在社群中的視覺人類簽章認證

An Evaluation of Visual Human Signature Identification in Community via Wearable Device

本論文係曹嘉慶君（學號 R01944003）在國立臺灣大學資訊網路與多媒體研究所完成之碩士學位論文，於民國一百零三年七月廿二日承下列考試委員審查通過及口試及格，特此證明

口試委員：

（簽名）

（指導教授）

陳祝嵩　　陳文進

所　長：

誌謝

隨著論文逐步的完成，我的碩士生涯也即將告一個段落。如果沒有家人、師長、同學、朋友的鼓勵與支持，我一個人一定無法完成這份研究。

首先我要感謝我的爸媽和我的兩個姐姐，他們在我心情低落時提供一個溫暖的避風港，在我開心之時，和我一起分享喜悅。感謝他們，讓我能夠在沒有後顧之憂的情況下致力於研究。

再來我要感謝我的指導教授徐宏民老師，這些日子以來，老師在研究上給予我許多的關懷與體諒，並在我遇到困難時，適時的提供專業的建議並點出盲點與實驗的缺陷，讓我在研究的道路上不至於迷失方向，知道下一步該做什麼；除此之外，老師的幽默風趣時常帶來笑果，讓開會時的討論不會過於苦悶。因為老師的這些幫助，讓這份研究得已完成。

接著要感謝實驗室中學長姐的鼓勵與支持，特別是 Yanying 學姐和 BorChun 學長的指導。Yanying 學姐時常與我一起討論研究上的細節，檢查實驗上的問題與發現問題的癥結；BorChun 學長會在我遇到瓶頸之時，提供突破性的思考，讓我能夠不繼續鑽牛角尖，找到新的方向。

還有要感謝所有 MiRAers，不管是同屆的、學長姐或是學弟妹，我們時常在研究室裡一起奮鬥、聊天說笑。一起出去玩時，能夠互相照顧，讓整個研究所生活變得多采多姿。

最後要感謝所以在我低落的時候陪我聊天的人，包括了糰 *13、球場上的朋友、大學同學。有了你們的陪伴，在我壓力大時幫我紓解壓力、渡過難關也讓我的研究得以順利完成。

Acknowledgements

I would like to express my deepest appreciation to my family, especially to my parents and my two sisters, for their continuous support during my master life. My parents always provide me the warmest environment to do research and study. Without their support, I will never sustain the pressure and overcome all the difficulties. I would like to express my greatest gratitude to my advisor, Prof. Winston H. Hsu, who is a role model in academic research. I have learned a lot from his inspirations and instructions since I joined the Communication and Multimedia Laboratory (CMLab) of the National Taiwan University. Without his encouragement and guidance, I would not have finished this thesis. I also want to express my appreciation to all my friends. Special thanks first go to my team leader, Yan-Ying Chen, for her mentoring, guidance and assistance from all aspects. Special thanks also go to all the MiRAers in my team, for cheering me up when I felt depressed and down. I also thank to my college friends and those who have played volleyball with me. They help me release the pressure and get over the hardest moment. I shall thank all my friends who have helped me for the past two years of my master life.

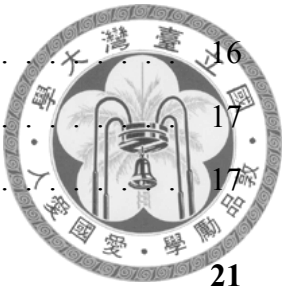
摘要

隨著穿戴式裝置的流行，我們逐漸能在不同的情境下接收到各種資訊。然而，囿於隱私權的規範，個人的資訊分享仍然是一個須要解決的問題。我們提出一種不管從哪個方向/姿勢都能表達一個人的概念–人類視覺簽章 (VHS)。使用者可以透過 VHS 將資訊散播於公開或是特定的社群中而不用顯示他們的身分。相對地，在社群中的人可以得知這些消息而不用知道這些人是誰。這篇論文探討了一些可能對於不同角度跟姿勢具有不變性的樣式來建造 VHS。我們評測諸多有效於在不同角度辨識人的樣式在不同角度跟姿勢的情況下的表現，樣式包含了人的臉部外觀、視覺方塊、臉部屬性、衣服屬性。我們還提出兩種用來融合多種樣式的方法 –提高重要的維度以及加權融合，前者用來增加召回率後者用來增加準確率。藉由同時考慮不同的樣式，我們提出的方法可以讓正面的 VHS 辨識達到 51% 的辨識率，在最難的測試集底下達到 23% 的辨識率。為了完整的評測我們的成果，我們介紹一個包含從許多人從不同角度觀測以及不同姿勢拍攝的全新資料庫 – 多角度名人個體資料庫 (MCID). 在這資料庫中擁有 439 位名人總共多於 2000 張從不同角度、不同服裝清晰的照片。據我們所知，這是截至目前為止能取得的資料庫中最大的。

Abstract

With the increasing popularity of wearable devices, information is becoming easily available anywhere and anytime. However, personal information sharing still poses great challenges because of privacy issues. We propose an idea of Visual Human Signature (VHS) which can represent each person uniquely even captured in different views/poses by wearable camera. Users can post information to certain communities or public by their VHS without reveal their identification. Conversely, the community can find the information while detecting the corresponding VHS via wearable devices. The thesis explores some possible modalities to generate VHS invariant to different views and different poses. We evaluate the performance of multiple modalities including person's facial appearance, visual patches, facial attributes and clothing attributes which are effective for recognizing identity in different views. We also propose two methods to fuse the modalities – emphasizing significant dimensions and weighted fusion; the former can improve the recall and the latter improve the precision. By jointly considering multiple modalities, our approach can achieve VHS recognition rate by 51% in frontal images and 23% in the most difficult dataset. To thoroughly evaluate our work, we introduce a new dataset for scenario of different view and clothing human retrieval called Multiview Celebrity Identity Dataset (MCID). The dataset contains more than 2,000 clarity images of 439 celebrities collected from web with different views and clothing. To the best of our knowledge, it is by far the largest publicly available multi-view and clothing dataset with identities.

Contents



List of Figures

List of Tables

Chapter 1

Introduction

In recent years, carrying wearable displays and cameras, such as camera-embedded glasses, becomes a trend. Who you see and what you confront can be sent to the server by the devices and bring a new vision in your life. The emerging technology poses a great opportunity to the share and grab information on the fly. However, users may not like to reveal too much about their identity while sharing information with the others. In this paper, our idea is to generate Visual Human Signature (VHS) from user's profile photo to represent themselves and share information with communities in the vicinity. Taking Figure 1.1 as an example, a user can attach a message to his/her VHS, saying "I am searching for a person to take taxi together." . The other users nearby can get this message if their wearable devices detect the message owner's VHS.

The problem is similar to human identification problem via the wearable device,which may confront the privacy issue that pedestrian admits to identify who he/she is. Nonetheless, the idea of VHS can solve this problem because the recognition is based on VHS. VHS can also distinguish people in the community which would like to exchange information in a more private way. Through VHS, users can represent themselves with an unique visual signature and share information without showing their identity. Meanwhile the signature could be easily updated if users upload new profile photos afterwards.

However, it is difficult to represent a target person with an unique signature by solely relying on a single modality, like face appearance or clothing features. In this work, we evaluate the performance and the limitations of different visual feature modalities for gen-

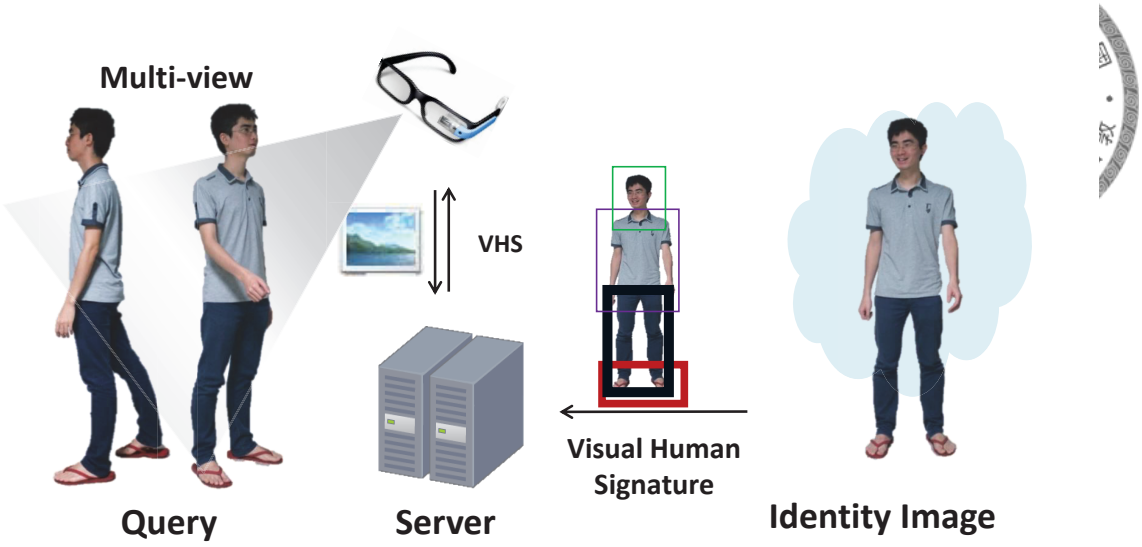


Figure 1.1: We propose to generate Visual Human Signature (VHS) as an unique representation of a target person even his/her image is captured in unconstrained environment via wearable devices. Users can leverage VHS to share information (e.g., a message for finding taxi-sharing partners) to the communities nearby once their wearable devices detect the message owner's VHS.

erating VHS. Although the meta-data helps image annotation and performs well [3], but in the real world, some meta-data is hard to collect. Hence, we only focus on the modalities related to visual features directly extracted from image content.

Four major visual modalities are considered and compared in this work, including facial appearance, visual patches, facial and clothing attributes. Facial appearance is the first modality coming to our mind when we aim at constructing VHS. Face is the most informative cue to find a target person. The power of facial appearance in human identification problem has been surveyed in [13]. Here we extract low-level feature on face landmarks to construct VHS. Another modality we considering is visual patches. Visual patches has been shown promising for scene classification [14]. Differently, we aim to find the significant patch in the identity image; for example, specific accessories or tattoo on body. Besides, we use facial attributes [10], which shows great impact in [9] work. We also jointly use clothing attributes which can highlight the difference between clothing

styles.[7] These four visual modalities though have limitations in certain circumstances, they are complementary to each other. We propose to firstly emphasize significant dimensions in each modality then jointly exploit multiple modalities to generated VHS for a target person and demonstrate the proposed VHS are more robust to view variations and can reach better accuracy compared to leaning on any of the single modalities. To sum up, our contributions include proposing the idea of visual human signature generation for sharing information in communities via wearable camera, discussing the challenges in different modalities and further improving the performance by emphasizing the significant dimensions and weighted voting methods.

Chapter 2

Related work

In some aspects, this problem is similar to image annotation for human [1, 15, 17]. In [1, 17], both use the contextual information to help the annotation on people. Their experiments show the importance of leveraging facial and clothing information in human image annotation. However, these methods are based on the people who can be detected by the face detection, which will failure when the picture is taken from the side or back, while the occlusion of faces often happens especially in the images freely taken by wearable devices.

Without using face information, Wang et al. [16] try to solve the human identification problem by capturing image from chest to head and generate the signature by extracting upper-body wavelets and spatiogram features. However, the features are not robust to pose and view variations for the target person.

Leyvand et al. [11] have proposed a similar idea through Kinect. They proposed to construct signature from

Chapter 3

Dataset Collection

Figure 3.1: The illustration of Multiview Celebrity Identity Dataset (MCID). MCID contains more than 2,000 clarity images of 439 celebrities with different views and clothing collected from web. To the best of our knowledge, it is by far the largest publicly available multi-view and clothing dataset with identities.

In this work, we need clearify identity's frontal and profile images with different clothes while there is no appropriate public dataset for us. So, we have crawled celebrity images from the webs as a dataset for our experiments and we name it Multiview Celebrity Identity Dataset (MCID). Overall, MCID contains 2341 clarity images of 439 popular

identities. Like the name of the dataset, each identity contains at least 2 images with different views and clothings. Images in this dataset have sufficient resolutions (about 500×750) for different modalities' need, such as pose estimator or attribute detection. To the best of our knowledge, it is by far the largest publicly available multi-view and clothing dataset with identities.

Chapter 4

Method

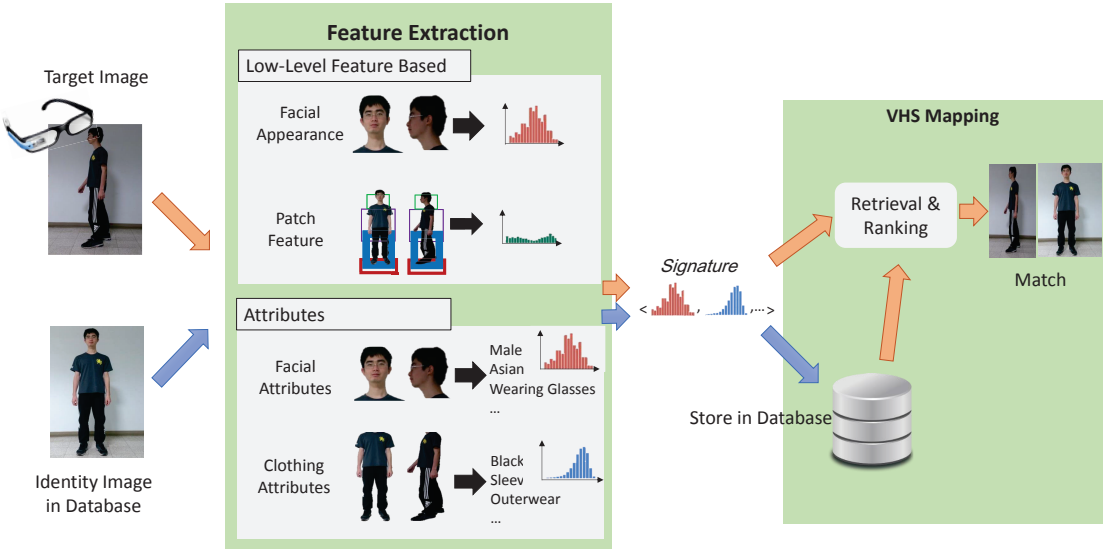


Figure 4.1: The proposed system. Four modalities of visual features are generated and jointly considered as the VHS of the target person from the uploaded frontal full-length image. Once the wearable devices upload any target person's image, our system will search for the most similar VHS.

As shown in Figure 4.1, the overall system comprises two phase. 1) User uploads his/her own frontal-view full-length image as profile image. With the image, our system constructs VHS by extracting the multi-model features and then keep the VHS in the database. The VHS of X can be noted as:

$$VHS_X =< M_1, M_2, ..., M_N > \tag{4.1}$$

where the M_i is the VHS generated from the ith modality. In this paper, we use facial appearance features, significant visual patches features, facial attributes and clothing attributes. 2) Given a target image captured by wearable device, we apply it with the same feature extraction process and generate VHS accordingly. We then compute the cosine similarity between the VHS in database in different modalities. Finally, we fusion the result in each modality and output the most similar VHS.

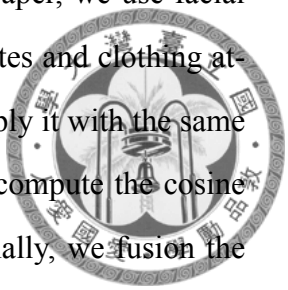


Table 4.1: The madalities we use to construct VHS. We use different kinds of low-level and mid-level features representing identity's facial information and clothing information. The number in the bracket following the attributes means how many classifiers/labels in the attributes.

Low-Level Features	Visual Patch Features	Codeword histogram (Trained from Lab color space histogram, HoG)	
	Facial Appearance Features	High-dimensional LBP	
Mid-Level Features		Binary Classifier	Multi-Label Classifier
	Facial Attributes	Race (3), Gender (2), Age (3), Glass	
	Clothing Attributes	Color(11), Patterns (6), Skin exposure, Scarf, Placket, Necktie, Gender, Collar	Category (8), Sleeve length (3), Neckline (3)

4.1 Visual Modalities

In this section, we introduce the modalities for generating VHS from user uploaded image. The total features we use are shown in Table 4.1.

4.1.1 Facial Appearance (FA)

Facial appearance is an informative cue in finding a target person. We first detect facial landmarks by face detection, including eyes, nose and mouth. Low-level features are extracted around each landmark by gridding into patches. Comparing the performance in different kinds of low-level features, we finally choose high dimensional local binary

pattern (HD-LBP) [4] which performs better than any other features to represent facial appearance.

4.1.2 Significant Visual Patch (PF)

To find the significant patch for each identity, we first detect human body in image content and divide it into patches. Then we extract features in patches and disallow highly overlapping patches or patches with low gradient energy. Without losing the generality, we consider only the square patches now and choose the size of patch range from 80x80 pixels to height of image size. We extract two kinds of features from each patch including:

- **Color–** Each patch is divided into 8x8 cells and extract the color features in LAB color space in each cell. To avoid illumination variation, we only use the mean value in A and B dimensions to represent each cell.

- **Histogram of Gradient (HoG) –** HoG is a well-known feature that can deal with the object detection problem and handle the texure details in image content. Here we generate the HoG descriptor in 8x8x31 cells with a stride of 8 pixels per cell [6].

Consequently, each patches has $8 \times 8 \times 31 + 8 \times 8 \times 2 = 2112$ dimensions. Afterwards, we adopt the Bag-of-Word model, which is a general model in representing an image. The features extracted from clothing dataset [5] 'are used in training a codebook with 512 dimensions. Finally, a 512-dim VHS is generated from the histogram of the clustered patch' features as the representation of the given image.

4.1.3 Attributes

Here we choose two categories of attributes to generate the VHS, facial attributes and clothing attributes.

- **Facial Attributes. (FAttr)**

  We utilize nine facial attributes in [9], including two gender attributes (female, male), four age attributes (kid, teen, middle-aged, elder) and three race attributes (Caucasian, African, Asian) to represent the different identities. Currently, we only

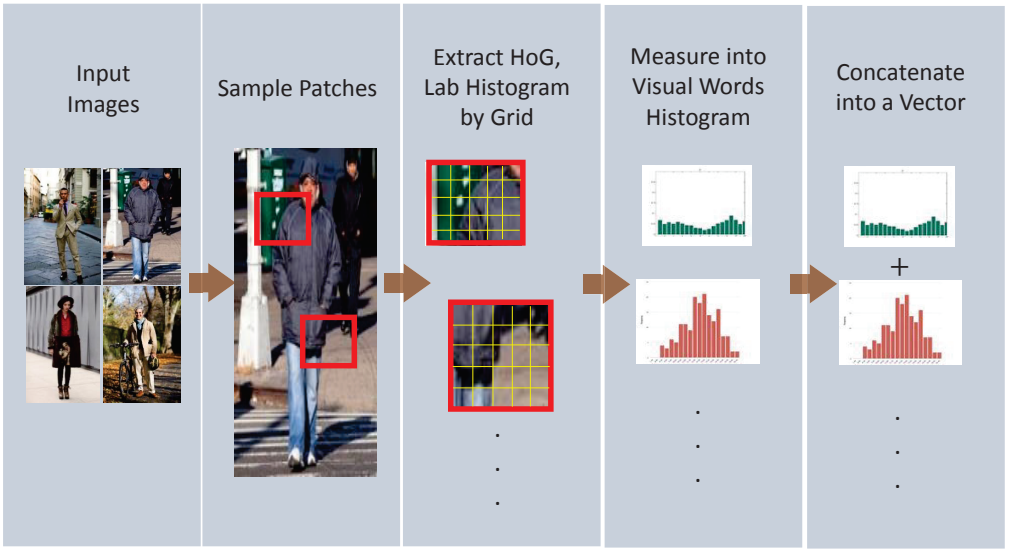


Figure 4.2: The framework of extracting significant visual patches. We first divide the image into patches and extract HoG and Lab histogram in each patch by grid. Then we measure each patch feature into visual word histogram by pre-trained codebook trained from same kinds of feature and procedure. At last, concatenate each codeword histogram into a vector as the feature of significant visual patch.

focus on the facial attributes since they represent rich information of people and can be learned through an adaptive framework [9]. The training dataset for facial attributes is collected from the Flickr. We extract Pyramid Histogram of Oriented Gradients (PHoG), Log-Gabor [12], Local Binary Patterns and Grid Color Moment in four face components (eyes, nose, mouth and whole face) from each image. To describe varying facial attributes, the classifier of each attribute is the most effective combination of regional representation trained by SVM and selected by Adaboost.

- **Clothing Attributes. (CAttr)**

  We use the clothing attributes defined in [5] to help generating VHS for a person. They define 26 clothing attributes, including 23 binary attributes (6 for clothing pattern, 11 for color and 6 miscellaneous attributes) and 3 multi-class attributes (sleeve

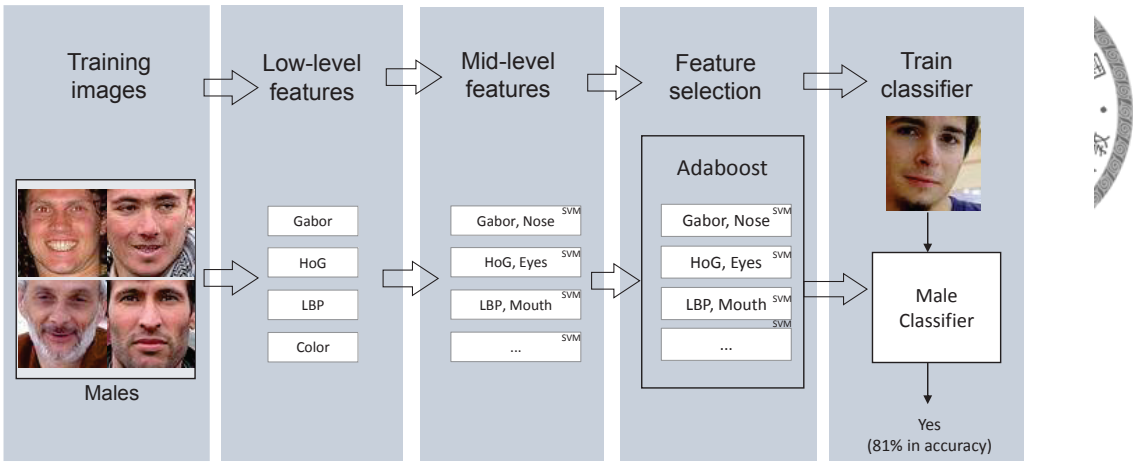


Figure 4.3: The framework of facial attributes. Process images with face detector to get bounding box and facial landmarks; then extract four kinds of features on each part, train mid-level SVMs and aggregate with Adaboost to form strong attribute classifier.

length, neckline shape and clothing category). Notice that the clothing attributes also include gender but it is designed to describe the clothing stylish, not the same as the facial attributes.

Each model is trained in the dataset provided from [5]. Most images in the dataset are pedestrians on the street. We duplicate the frameworks in [5], extracting 40 features, performing SVM classification by combined features, employing the Conditional Random Field to learn relationships between the attributes.

The attribute detector outputs a vector of probability for each binary attributes label. These probabilities reflect the confidence of the attribute prediction. The VHS of attributes is then generated by concatenate the probability from each detector.

4.2 Signature Matching and Modality Fusion

After extracting the features, we want to calculate the similarity between target's VHS and VHS in database and rank by the similarity. Here, we apply two strategies to improve the final performance – emphasizing significant dimensions and weighted voting. The former can improve the recall in each modality by deweighting the meaningless dimensions and

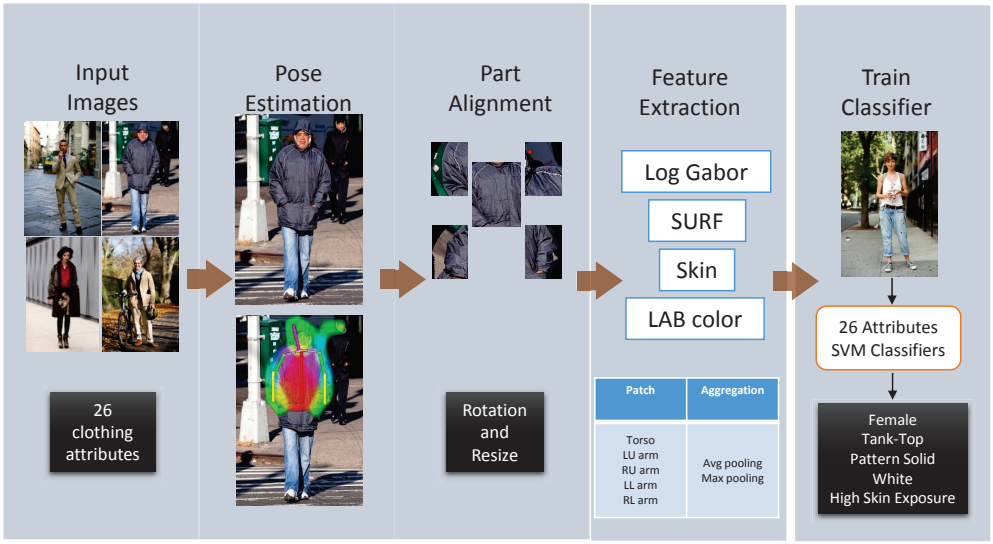


Figure 4.4: The framework of learning clothing attributes. First, using pose estimator to detect torso, arms and legs. Second, extract 40 features in each segment; then, perform SVM classification by combined features. At the end, employ the Conditional Random Field to learn relationships between the attributes.

weighting the significant dimensions while the latter can improve the precision by integrating the similarity of different methods which shows the effect of tolerating missing information in different modalities.

4.2.1 Emphasize Significant Dimensions

In this step, we want to hightlight the informative dimensions in each modality. Since our dataset is composed of a certain community, some attributes' possibility response are similar or some clothing attributes can bring more information. For example, the score of race attribute might be less informative and should be deweight in an Asian dataset, . To achieve the goal, we calculate the mutual information between the identity in the dataset in each modality. The equation can be written as

$$MI(X;Y) = H(X) - H(X|Y) = \sum_{x \in X} p(x) \times \log p(x) - \sum_{x \in X} \sum_{y \in Y} p(x,y) \times \log \frac{p(y)}{p(x,y)} \tag{4.2}$$

,where X denotes the identity and the Y represent a modality's dimension. As shown in 4.3, the similarity of one modality is then computed with weighting in each dimensions

by the normalized mutual information of each dimension, where p is the VHS in dataset and q is the VHS of input image. $v_{p,i}$ and $v_{q,i}$ is the the value of p and q at dimension i, and MI_i is the normalized mutual information.

$$Sim(p,q) = \frac{\sum_{i=1...D} MI_i \cdot v_{p,i} \cdot v_{q,i}}{|p| \cdot |q|} \tag{4.3}$$

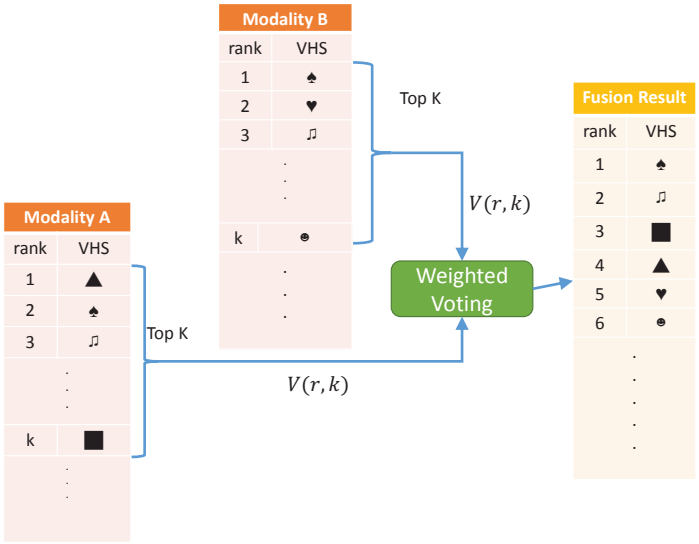


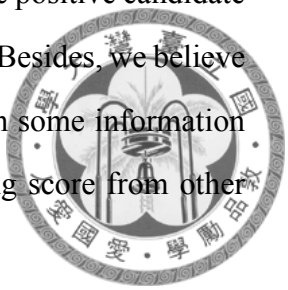
Figure 4.5: The weighted fusion of each modality. We give top-K candidate a score V in each modality. Later, re-rank the candidate VHS by the summation of voting score V in each modality.

4.2.2 Weighted Voting

After ranking the VHS by the similarity, we want to levarage the ranking list from each modality. As shown in Figure 4.5, for a VHS ranking r, a score V is given in each list from top K candidates. The function of score V can be written as:

$$V(r,k) = \begin{cases} k+1-r & \text{if } r \leq k \\ 0 & \text{if } r > k \end{cases} \tag{4.4}$$

Only top K can get the score because, in our scenario, there is only one positive candidate in the dataset. The larger K we choose, the more noise will be chosen. Besides, we believe the proposed method can tolerate some missing information. Though some information (e.g., facial attributes) might be missing, we can still get high voting score from other modalities (e.g., clothing attributes).



Chapter 5

Experiment

5.1 Experiment Settings

To evaluate our work, we generate 4 image sets from 300 identities in MCID mentioned in Chapter 3.

- **DataSet –** There are 300 different identity's frontal full-length image in DataSet. Namely, each identity has one frontal full-length image.
- **FrontalSet –** FrontalSet is composed of 100 images. Each image contains a frontal-shot identitiy dressing the same as what he/she wears in Dataset.
- **ProfileSet –** ProfileSet consists 100 images. Each image contains a profile-shot identitiy dressing the same as what he/she wears in Dataset.
- **AllSet –** AllSet is made up of 1309 images. Each image contains an identity that is possibly shot in different view or dressed in different clothing styles.

We use DataSet to construct the identity VHS and regard it as our dataset; the others are used as our testset. Notice that AllSet is the most difficult testset in these testsets because it consists of multiview identities' images and identities in different clothing.

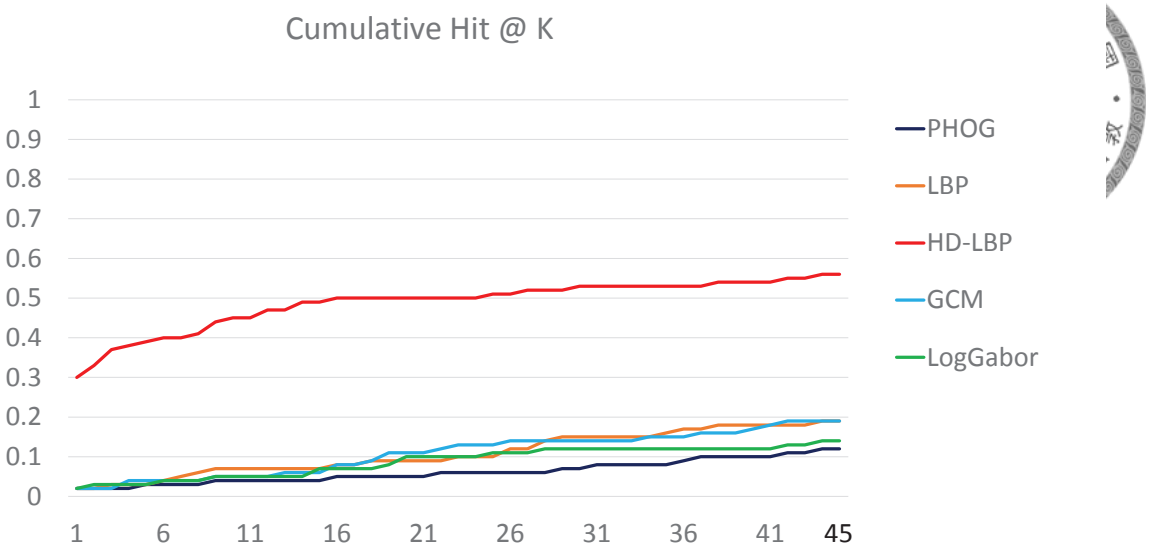


Figure 5.1: The performance of facial appearance. It is obvious that HD-LBP [4] overwhelms other low-level features in FrontalSet by reaching Cumulate Hit = 0.3 at rank 1. But all of the curves climb slowly as the rank grows.

5.2 Performance Evaluation

5.2.1 Evaluation

We here use the Cumulative Hit at K curve suggested in [8]. The curve represents the cumulated values of recognition rate at all ranks. The value is computed as:

$$Cumulative\,Hit\,@\,K = \frac{\sum_{i=1}^{N}\sum_{j=1}^{K} H(j)}{N}, \quad H(j) = \begin{cases} 1 & \text{,if hit} \\ 0 & \text{,otherwise} \end{cases} \tag{5.1}$$

,where N is the total query number.

5.2.2 Different Features in Facial Appearance

Using FrontalSet as testset, we evaluate the performance of different kinds of low-level features in facial appearance . We have extracted Pyramid HoG (PHOG), Log-Gabor, Grid Color Moment (GCM), Local Binary Pattern(LBP) and High Dimensional Local Binary Pattern (HD-LBP). As shown in Figure 5.1, it is obvious that HD-LBP [4] overwhelms

other low-level features in FrontalSet by reaching Cumulate Hit = 0.3 at rank 1. The reason is that HD-LBP extracts more information by down/up scaling and uses more dimensions to describe the details.

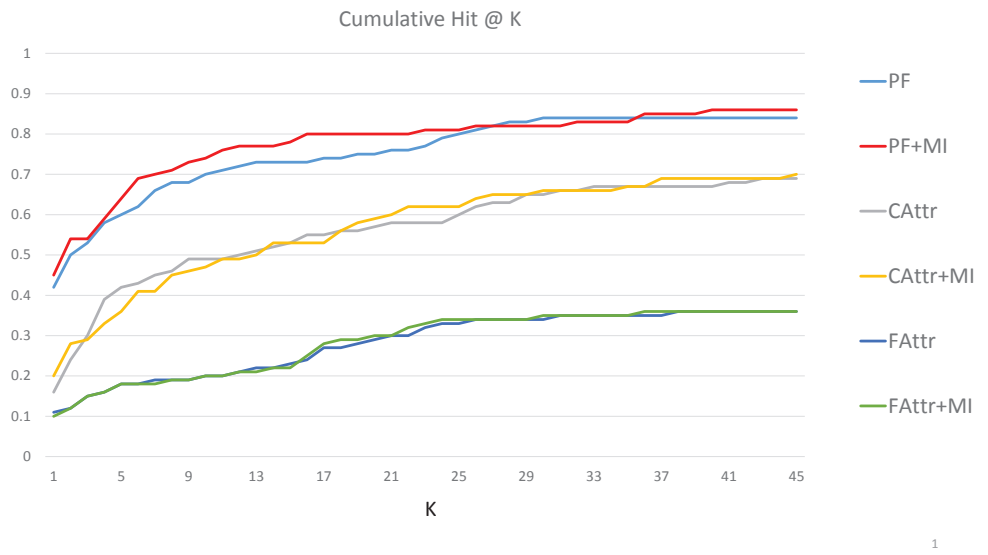


Figure 5.2: The performance of emphasizing significant dimensions. The performance has been improved about 0.1 in PF. We can see the mothod has improved the recall rate in all modalities.

5.2.3 Gain of Emphasize Significant Dimensions

Tested by FrontalSet, emphasizing significant dimensions is performed in PF, FAttr and CAttr. The method is not used in FA because feature we used in FA is not suitable for the method. The result is shown in Figure 5.2. The recall rate has been improved about 0.1 at K=15 in PF while in other modalities only improved about 0.03. We think the reason is the dimensions in these two modalities are not enough to highlight the significant dimensions.

5.2.4 Performance in Different Testset

We test our frameworks in FrontalSet, ProfileSet and AllSet. We also show the performance of each modality after emphasizing significant dimensions. To compare with our weighted voting method, the performance of average fusion method is performed as well.

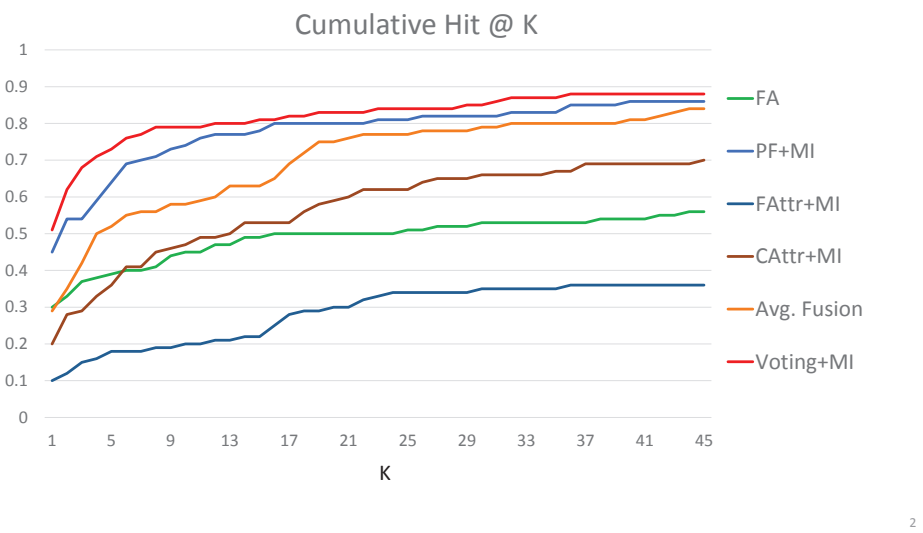


Figure 5.3: The performance of testing in FrontalSet. PF performs the best over all modalities on 0.46 at K=1. After weighted voting, performance improves 0.1 than Avg. fusion.

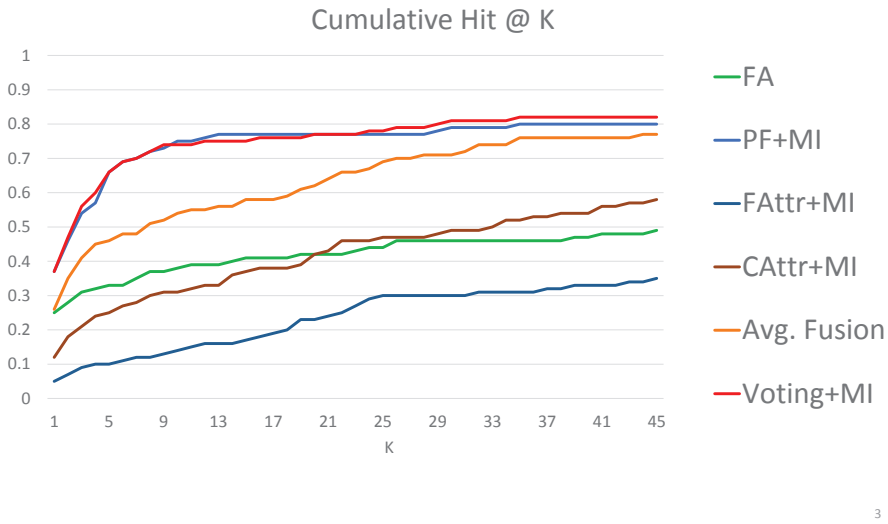
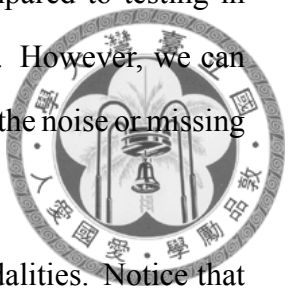


Figure 5.4: The performance of testing in ProfileSet. We can see the performance drop comared to 5.3 caused by the losing of facial information. But, the weighted voting keeps the better ranking and tolerate the noise or missing information.

- **FrontalSet:** Figure 5.3 shows the performance of our methods in FrontalSet. We can discover the significant visual patch features performs the best over all modalities. After weighted voting, performance improves 0.1 than Avg. fusion.

- **ProfileSet:** The performance is shown in Figure 5.4. Compared to testing in FrontalSet, all modalities' performance drop about 0.1 at K=1. However, we can see that the weighted voting keeps the better ranking and tolerate the noise or missing information.



- **AllSet:** As shown in Figure 5.5, PF performs the best in modalities. Notice that CAttr fails in AllSet because of different clothing. Still, after weighted voting, the performance climbs to 0.22 at K=1 while the Avg. fusion performs 0.16.

We have shown examples of tolerating the noise or missing information in Figure 5.6. Our proposed method can find the right identity images even the identity wears different clothing. In 5.6 (a), identity image for query is taken in profile; in other words, we loss the facial information. But positive candidate appears at rank 4 ascribed to PF and CAttr. As in 5.6 (b), the query identity wears different from what she wears in dataset's VHS. However, result shows the positive candidate's VHS is the most similar in the database. The reason is the facial information is clearify and strong enough in weighted voting.

Table 5.1: Cumulative Hit @ K of different modalities over images in AllSet. Adding weighted voting fusion achieves Cumulateive Hit of 0.48 at K=10, which outperforms 0.16 than average fusion. The performance of clothing attributes becomes very poor because of different clothing while face appearance still has Cumulative Hit of 0.21 at K=1.

	FA	PF	FAttr	CAttr	Avg. Fusion	Voting Fusion
K=1	0.21	0.26	0.06	<0.01	0.16	0.23(↑0.07)
K=10	0.31	0.43	0.15	0.06	0.32	0.48(↑0.16)
K=25	0.45	0.50	0.26	0.14	0.43	0.56(↑0.13)

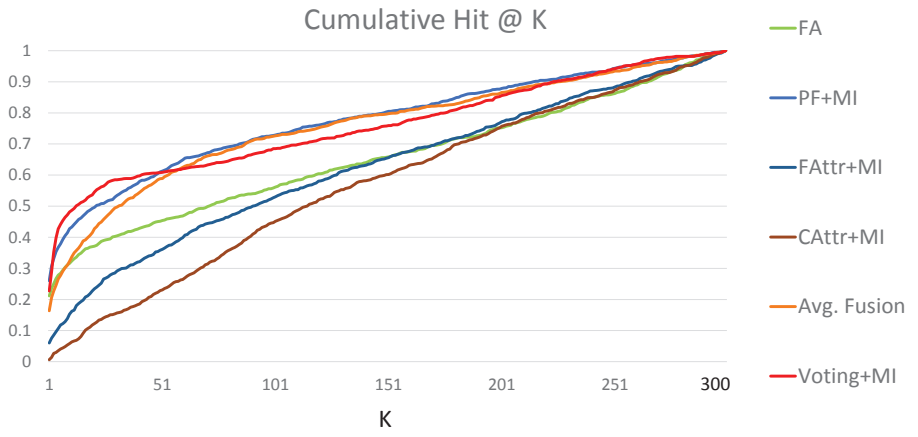


Figure 5.5: The performance of testing in AllSet. Notice that CAttr fails in AllSet because of different clothing. Still, after weighted voting, the performance climbs to 0.22 at K=1.

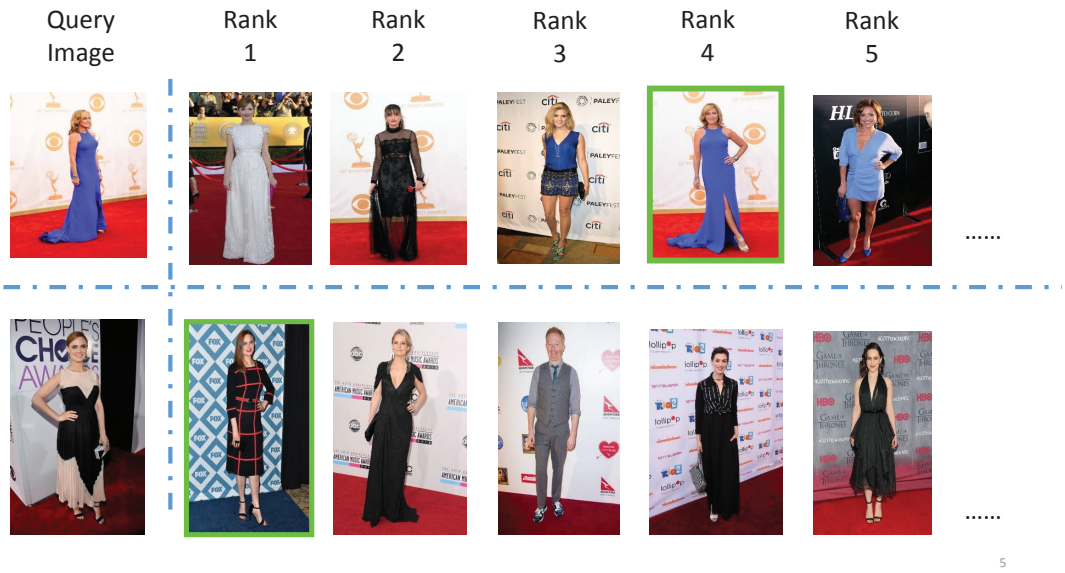


Figure 5.6: Ranking result. In (a), the query is a profile image missing the facial information, but we can find one with similar dressing, i.e. blue dress. In (b), the identity wears a totally different cloth from what she wears in the VHS. But we map the VHS by the frontal facial information. Both cases show our porposed method can tolerate some missing information.

Chapter 6

Conclusion

In this thesis, we discuss the challenge of generating Visual Human Signature which can be detected in unconstrained environment via camera-embedded wearable devices. We collect a dataset named Multiview Celebrity Identity Dataset (MCID) containing 2341 images of 439 celebrities with different views and clothing. We also propose the idea of VHS, compare the performance of different modalities and show a preliminary evaluation in MCID by fusing the modalities in this new coming problem. Two methods, emphasizing significant dimensions and weighted voting, are employed in this thesis to solve and improve the performance. The results encourage many directions in which this work can be extended. For example, this problem can be further extended to taking a video as an input target. Additional cues can also be incorporated, such as time stamps which can be easily fetched from the wearable devices. Or, with GPS information, we can scale down the searching area to improve the precision. We believe this people-centric sensing problem [2] will become more important, interesting and be get more attention in the future.

Bibliography

[1] D. Anguelov, K. chih Lee, S. Gokturk, and B. Sumengen. Contextual identity recognition in personal photo albums. In *CVPR*, 2007.

[2] A. Campbell, S. Eisenman, N. Lane, E. Miluzzo, R. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, and G.-S. Ahn. The rise of people-centric sensing. *Internet Computing, IEEE*, 12:12–21, July 2008.

[3] L. Cao, J. Luo, and T. S. Huang. Annotating photo collections by label propagation according to multiple similarity cues. In *ACM Multimedia*, 2008.

[4] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimisionality: High dimensional feature and its efficient compression for face verification. 2013.

[5] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, 2012.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[7] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *CVPR*, 2008.

[8] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro*, 2007.

[9] N. Kumar, P. N. Belhumeur, and S. K. Nayar. Facetracer: A search engine for large collections of images with faces. In *ECCV*, 2008.

[10] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *PAMI*, 33:1962–1977, 2011.

[11] T. Leyvand, C. Meekhof, Y.-C. Wei, J. Sun, and B. Guo. Kinect identity: Technology and experience. *Computer*, 2011.

[12] J. Li, T. Wang, and Y. Zhang. Face recognition using feature of integral gabor-haar transformation. In *ICIP*, 2007.

[13] A. Samal and P. A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recogn.*, 1992.

[14] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[15] D. Wang, S. Hoi, Y. He, J. Zhu, T. Mei, and J. Luo. Retrieval-based face annotation by weak label regularized local coordinate coding. *PAMI*, 36:550–563, 2014.

[16] H. Wang, X. Bao, R. R. Choudhury, and S. Nelakuditi. Insight: Recognizing humans without face recognition. In *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*, 2013.

[17] L. Zhang, L. Chen, M. Li, and H. Zhang. Automated annotation of human faces in family albums. In *ACM Multimedia*, 2003.