IDENTIFY VISUAL HUMAN SIGNATURE IN COMMUNITY VIA WEARABLE CAMERA

Chia-Chin Tsao, Yan-Ying Chen, Yu-Lin Hou, Winston H. Hsu

National Taiwan University, Taipei, Taiwan

ABSTRACT

With the increasing popularity of wearable devices, information becomes much easily available. 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 cameras. We evaluate the performance of multiple effective modalities for recognizing an identity, including facial appearance, visual patches, facial attributes and clothing attributes. We propose to emphasize significant dimensions and do weighted voting fusion for incorporating the modalities to improve the VHS recognition. By jointly considering multiple modalities, the VHS recognition rate can reach by 51% in frontal images and 48% in the more challenging environment and our approach can surpass the baseline with average fusion by 25% and 16%. We also introduce Multiview Celebrity Identity Dataset (MCID), a new dataset containing hundreds of identities with different view and clothing for comprehensive evaluation.

Index Terms— Visual Human Signature, Human Attributes, Wearable Device, Weighted Voting

1. INTRODUCTION

In recent years, wearable displays and cameras, such as camera-embedded glasses, have become a trend. Who and what you confront can be sent to the server by the devices and bring a new vision in your life. This emerging technology poses a great opportunity to share and grab information on the fly. However, users may not like to reveal their identity while sharing information. In this paper, our idea is to generate Visual Human Signature (VHS) from user's frontal photo to represent themselves and share messages with communities in the vicinity. As shown in Figure 1, a user can attach augmented information to his/her VHS, and once other users' wearable devices nearby detect the VHS, they can get the information that the owner of VHS wants to share; for example, one shares the information of finding the partner of two-for-one offer via VHS.

VHS is a form to represent a person for the purpose of information sharing without revealing his/her identity. Such nature makes it different from human identification which targets on recognizing a person's identity. [1, 2] Furthermore, the information sharing is constrained by locality such that users can keep their information private to the community nearby rather than make it public. Meanwhile, the signature could be easily updated if users upload new profile photos afterwards.

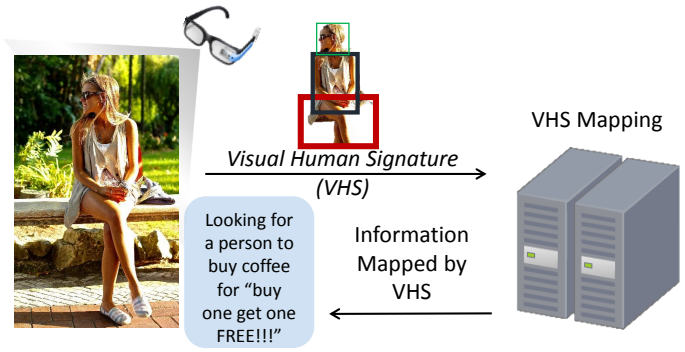

Fig. 1. The scenario of using Visual Human Signature (VHS). We propose to generate VHS as an unique representation of a target person even his/her image is captured in unconstrained environment via wearable devices. Users can leverage VHS to share information (e.g., a message for finding buy-one-get-one-free partner) to the communities nearby once others wearable devices detect the message owner's VHS.

Using facial and clothing information has been shown effective to differentiate persons [3, 4]. [3] proposes to train classifiers with facial and clothing features as the information in language model for each identity. [4] further exploits the facial and clothing cues with a Bayesian framework to distinguish people in images. These methods need training data with labels to learn the classifiers for each identity. However, keeping the unique VHS, we can distinguish people without other training data labeled with identities.

Wang et al. propose to use the clothing information to represent a person through mobile phones [5]. Without using facial information, they capture images from chest to head and generate the signature by upper-body wavelets and spatiogram features. But the features are not robust to solve pose and view variations for the target person. Leyvand et al. [6] also propose to construct signatures from clothing and depth information through Kinect; nonetheless, their work focus on distinguishing few identities in Kinect's applications and is limited to the shooting coverage of Kinect's camera while we propose to deal with more people in the local community.

In this paper, four major visual modalities, including facial appearance, significant visual patches feature, facial and clothing attributes, are considered and compared to reduce the

effects from the pose, view and clothing variations. Facial appearance is the most informative cue to find a target person. However, only relying on facial appearance is not enough because of occlusion and expression [7]. Therefore, we also consider visual patches extracted from whole body. Visual patch has been shown promising for scene classification [8]; however, we aim to find the significant patches in identity images; for example, specific accessories or tattoo on body. Besides, we use facial attributes [9], which shows a great impact on face retrieval [10]. We also jointly use clothing attributes which can highlight the difference between clothing styles [11].

Each modality has the problem of missing information in some circumstances, for example lacking of the facial information in profile view. To overcome this problem, we incorporate multiple modalities that are complementary to each other to generate VHS for a target person. Then, we calculate its similarity scores of any two VHS by firstly emphasizing significant dimensions in each modality and conducting the weighted voting. We demonstrate the proposed VHS are more robust to the varying views and can reach better accuracy compared to leaning on any of the single modalities.

To evaluate the capability for tackling the intra-class variations in viewing angles and clothing styles, we collect celebrity image set from the web, referred as Multi-view Celebrity Identity Dataset (MCID). Overall, MCID contains at least 2 images of each identity with different views and clothings styles. To the best of our knowledge, it is by far the largest publicly available dataset of human photos designed for studying the problem of multi-view and clothing styles.

In summary, our contributions include – (1) Proposing the idea of visual human signature for sharing/receiving information in communities via wearable cameras. (2) Discussing the challenges in different modalities and further improving the performance by emphasizing the significant dimensions and weighted voting methods. (3) Introducing a new multiview and clothing dataset, MCID, for evaluating VHS identification problem. The dataset contains 2,341 images with 439 identity labels and is made publicly available.

2. SYSTEM OVERVIEW

As shown in Figure 2, the overall system comprises two phases. 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. Then, given a target image captured by the wearable device, we apply it with the same feature extraction process and generate VHS accordingly. We then measure the similarity between each VHS in database and output the most similar VHS.

3. VISUAL HUMAN SIGNATURE (VHS)

We use following four modalities for generating VHS from user uploaded images. The VHS of attributes is generated by concatenating the features of each modality.

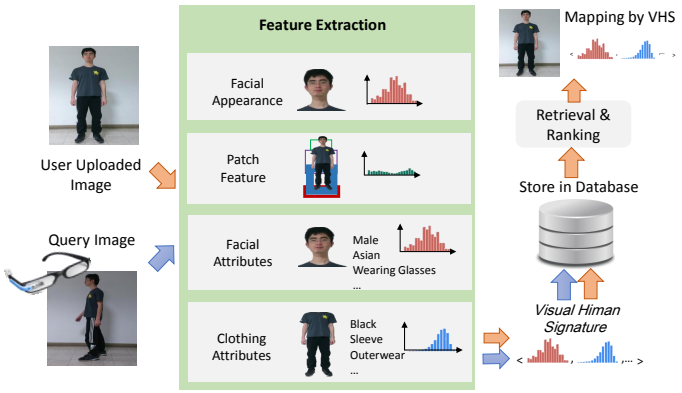

Fig. 2. The proposed system. Four modalities of visual features are generated and jointly considered as the VHS of the target person from the uploaded frontal full-length image. Once the wearable devices upload any target person's image, our system will search for the most similar VHS.

– **Facial Appearance (FA)**: Facial appearance is an informative cue for finding a target person. We firstly detect facial landmarks, including eyes, nose and mouth. Low-level features are then extracted around each landmark by grids. Comparing the performance in different kinds of low-level features, we finally choose high dimensional local binary pattern (HD-LBP) [12], which uses more dimensions to describe the details by up/down sampling at the landmarks. The small scale describes the detailed appearance around the fiducial points and the large scale captures the shape of face in relative large range.

– **Significant Visual Patch Feature (PF)**: To visually represent a person by the significant patches on his/her body, we extract features in patches of size ranging from 80x80 pixels to the whole image and reject highly overlapping patches or patches with low gradient energy. We extract two kinds of features from each patch including color histogram and HoG. Each patch is divided into 8x8 cells and extracts LAB color features, which only use the mean values in A and B channels for avoiding illumination variation. The HoG descriptor is generated in $8 \times 8 \times 31$ cells with a stride of 8 pixels per cell. Afterwards, we adopt the Bag-of-Word model, which is a general model in representing an image. The features extracted from clothing dataset [13] are used in training a codebook with 512 dimensions. Finally, a VHS is generated from the histogram vector of the codebook.

– **Facial Attributes (FAttr)**: We utilize nine facial attributes in [10], including two gender attributes (female, male), four age attributes (kid, teen, middle-aged, elder) and three race attributes (Caucasian, African, Asian) to represent a person. The training dataset for facial attributes is collected from the Flickr. We extract Pyramid Histogram of Oriented Gradients (PHoG), Log-Gabor [14], Local Binary Patterns (LBP) and Grid Color Moment (GCM) 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 [13] as one of the modalities to represent a person. Totally, 26 clothing attributes, including 23 binary attributes (6 for clothing pattern, 11 for color and 6 miscellaneous attributes) and 3 multi-class attributes (sleeve length, neckline shape and clothing category), are learned to describe clothing styles. Each attribute classifier is trained in the dataset provided from [13] where most images comprise pedestrians on the street. We duplicate the frameworks in [13], extracting 4 low-level features from 5 body parts (torso, two upper arms, two under arms) and performing SVM classification by combining features in max-pooling and average-pooling. Finally, 59 dimensions are used from clothing attribute classifiers response to describe a person.

4. VHS MATCHING AND MODALITY FUSION

After extracting the features, we propose to calculate the similarity between the target person's VHS and the VHS in the database. To incorporate result of different modalities, we propose two strategies for fusion– emphasizing significant dimensions and conducting weighted voting. The former can improve the recall in each modality by highlighting the significant dimensions, and the latter improves the precision by integrating the similarity of different modalities which reduces the effects of missing information in single modalities.

4.1. Emphasizing Significant Dimensions

We aim to highlight the informative dimensions in each modality. Since our VHS in the dataset are composed of a certain community, some of their attributes' responses are similar and present little information while of them are more discriminative and contribute much more information. For example, the scores of race attribute is less informative in an Asian community and the weight should be suppressed. To achieve the goal, we calculate the mutual information which represents how much uncertainty of differentiating persons X (in database) is reduced after knowing the values of a dimension Y in a modality.

$$MI(X;Y) = \sum_{x\in X} p(x)\times \log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\times \log \frac{p(y)}{p(x,y)} \tag{1}$$

, where X denotes the identities in the database and Y is the any one dimension of VHS. The similarity between the VHS p in the dataset and the VHS q from input image is calculated as:

$$Sim(p,q) = \frac{\sum_{i=1}^{|Y|} w_i \cdot v_{p,i} \cdot v_{q,i}}{|p|\cdot|q|} \tag{2}$$

, where $v_{p,i}$ and $v_{q,i}$ are the value of p and q VHS at dimension i, and w_i represents the weight for the i_{th} dimension, which is the normalized mutual information of i_{th} dimension.

4.2. Weighted Voting

To improve the precision, we rank the VHS in the dataset by the similarity and incorporate the ranking list of each modality. In each modality, only top T candidates, which are the most similar to the query's VHS, will receive a score V; otherwise, the discriminative power will be amortized. The V is defined as

$$V(r,T) = \begin{cases} T+1-r & \text{if } r \le T \\ 0 & \text{if } r > T \end{cases} \tag{3}$$

, where r is the rank. Though some information (e.g., facial attributes) might be missing, we can still obtain high voting score from other modalities (e.g., clothing attributes). In other words, the proposed method can tolerate some missing information.

5. EXPERIMENT

5.1. Multiview Celebrity Identity Dataset (MCID)

Because there is no appropriate public dataset for us to evaluate the robustness in pose and view variation of VHS, we collect the images from the website, celebrity-gossip.net, where contains lots of celebrities' images. An important criterion is concerned that the dataset should be made up of different view and pose. Hence, we pick up the celebrities who have full-length images with different clothing and in multi-view, totally 2,341 images of 439 identities. The resolution of each image in MCID is about 500×750.

To simulate the scenario of a community, we pick up 300 different identities from MCID to construct the VHS database with frontal images. Three sets are generated to test the performance in different situations – FrontalSet, ProfileSet and AllSet. FrontalSet is made up of 100 images containing only frontal view. ProfileSet consists 100 profile images and AllSet is composed of 1,309 images containing multiview of the identities. The identities in former two test sets are dressed the same as in the dataset, while the last may dress in different. Notice that AllSet is the most difficult test set because it consists of multiview identities' images and identities in different clothing.

Table 1. The statistic of Cumulative Hit with/without emphasizing significant dimensions. The details are explained in Section 5.2.

	PF	PF +MI	FAttr	FAttr +MI	CAttr	CAttr +MI
K=1	0.42	**0.45**	0.10	**0.11**	0.16	**0.20**
K=10	0.7	**0.74**	0.20	**0.20**	0.47	**0.49**

5.2. Performance Evaluation

– **Different Features in Facial Appearance**: The performance is evaluated through the Cumulative Hit curve suggested in [15]. The curve is the cumulated values of recognition rate at all ranks.

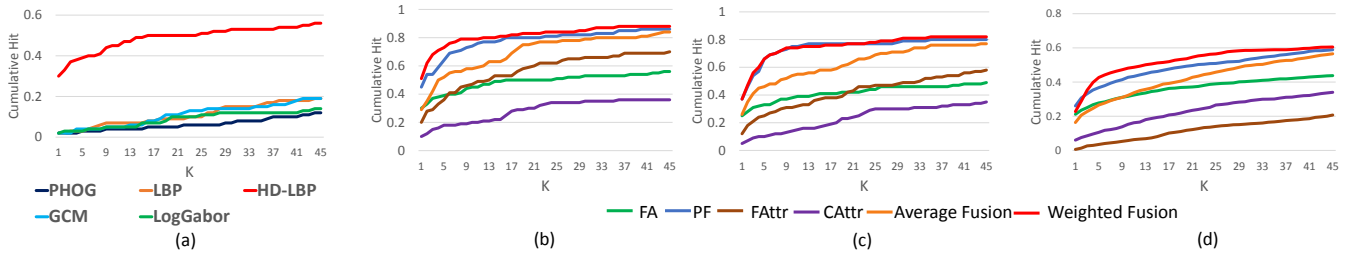

Fig. 3. (a) The performance of different low-level features describing facial appearance. HD-LBP [12] overwhelms other low-level features in FrontalSet. (b) The results of testing FrontalSet. The proposed Weighted Fusion outperforms any of the single modalities and the Average Fusion. (c)(d) The results of testing ProfileSet and AllSet. The weighted voting keeps the better Cumulative Hit which suggests it is more robust to noise or missing information. Though weighted fusion method is underperformed than PF initially, but outperformed others then.

Using FrontalSet as testset, we evaluate the performance of different kinds of low-level features in facial appearance, including PHoG, Log-Gabor, GCM, LBP and HD-LBP. As shown in Figure 3(a), HD-LBP overwhelms other low-level features and the CummulativeHit reaches 0.3 at K = 1. The higher performance may attribute to the much higher dimension of HD-LBP designed for describing the details with multi-scales.

– **Gain of Emphasizing Significant Dimensions**: Testing on FrontalSet, the result of emphasizing significant dimensions in modalities is shown in Table 1. All modalities' recall rate have improved for highlighting the informative dimensions. The higher recall rate is beneficial for the stage of weighted voting because it promotes the correct VHS to the top candidates.

– **Performance in Different Test Sets** We test our frameworks with $T = 10$ in weighted voting fusion method in FrontalSet, ProfileSet and AllSet. To compare with our proposed method, average fusion, averaging the similarity from different modalities, is performed as well. In Figure 3(b), PF performs the best over all modalities because PF includes not only the color information but also the details in the images. The weighted voting outperforms the average fusion in Cumulative Hit by 10%, which implies the modalities are not equally important and our proposed method can highlight the discriminative modalities. Because of the missing facial information and different views, the performance of all modalities' drop about 0.1 at K=1 ProfileSet as shown in 3(c). However, the weighted voting keeps the better ranking and is robust to the noise and missing information. Both FAttr and CAttr fail in AllSet because of different view and clothing, while PF keeps the best performance over all single modality, which is shown in 3(d). Still, after weighted voting, the performance climbs to 0.48 at K=10 surpassing the average fusion by 0.16.

– **Example Results** In this section, we visualize results using two example retrieval results as shown in Figure 4. Our proposed method can find the correct identity images even the identity's clothing style has been changed. In Figure 4(a), even we lose the frontal facial information but the proposed approach can still match the target person at rank 4 ascribed to PF and CAttr. In Figure 4(b), the result list shows the target person's VHS is at first rank in the database even if the query identity wears different from what she wears in her image in database. Both examples show that jointly considering multiple modalities can suppress the effects from information loss in different circumstances.

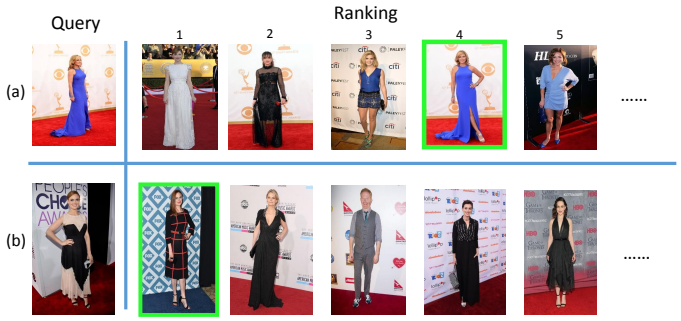

Fig. 4. Two examples of ranking result. In (a), the query is taken in profile but the target can be found at rank 4 by similar dressing. In (b), though the dresses in the query and the database images are different, the other modalities, e.g., face appearance, can help to match VHS. Both cases show our proposed method is considerably robust to missing information.

6. CONCLUSION AND FUTURE WORK

In this paper, we propose Visual Human Signature which can be used in unconstrained environment via camera-embedded wearable devices and share information in local communities. Two methods, emphasizing significant dimensions and weighted voting, are proposed to improve matching VHS under the variations of pose, view and clothing. We collect a dataset, MCID, which is the largest dataset with different views and clothing for comprehensive evaluations. The results encourage many directions in which this work can be extended. For example, additional cues can also be incorporated, such as time stamps which can be easily fetched from the wearable devices. Furthermore, with GPS information, we can scale down the searching area to improve the precision. We believe this people-centric sensing problem will become more important and get more attention in the future.

7. REFERENCES

[1] Amit Kale, N. Cuntoor, and R. Chellappa, "A framework for activity-specific human identification," in *ICASSP*, 2002.

[2] Tzu-Yi Hung, Jiwen Lu, Junlin Hu, Yap-Peng Tan, and Yongxin Ge, "Activity-based human identification," in *ICASSP*, 2013.

[3] D. Anguelov, Kuang chih Lee, S.B. Gokturk, and B. Sumengen, "Contextual identity recognition in personal photo albums," in *CVPR*, 2007.

[4] Lei Zhang, Longbin Chen, Mingjing Li, and Hongjiang Zhang, "Automated annotation of human faces in family albums," in *ACM Multimedia*, 2003.

[5] He Wang, Xuan Bao, Romit Roy Choudhury, and Srihari Nelakuditi, "Insight: Recognizing humans without face recognition," in *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*, 2013.

[6] T. Leyvand, C. Meekhof, Yi-Chen Wei, Jian Sun, and Baining Guo, "Kinect identity: Technology and experience," *Computer*, 2011.

[7] Ashok Samal and Prasana A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern Recogn.*, 1992.

[8] Saurabh Singh, Abhinav Gupta, and Alexei A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *ECCV*, 2012.

[9] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar, "Describable visual attributes for face verification and image search," *PAMI*, vol. 33, pp. 1962–1977, 2011.

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

[11] A.C. Gallagher and Tsuhan Chen, "Clothing cosegmentation for recognizing people," in *CVPR*, 2008.

[12] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun, "Blessing of dimisionality: High dimensional feature and its efficient compression for face verification," 2013.

[13] Huizhong Chen, Andrew Gallagher, and Bernd Girod, "Describing clothing by semantic attributes," in *ECCV*, 2012.

[14] Jianguo Li, Tao Wang, and Yimin Zhang, "Face recognition using feature of integral gabor-haar transformation," in *ICIP*, 2007.

[15] Doug Gray, Shane Brennan, and Hai 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.