Person Tagging in Still Images by Fusing Face and Full-body Detections

Vlastislav Dohnal and Alexander Matecny
Faculty of Informatics
Masaryk University
Brno, Czech Republic
Email: dohnal@fi.muni.cz / shanio@mail.muni.cz

***Abstract*—We address the problem of organizing personal photo albums by assigning tags/names to people present in photographs. Our proposed framework improves similar systems such as Google+ Photos (Picasa) or Apple's iPhoto by incorporating not only a face detector, but also a full-body detector. Both these modalities are combined together to provide the user with tags of people whose face has not been detected or is not even present in the photograph. An implementation of the proposed framework is evaluated on a sample of real life photographs. This paper is a "work in progress" contribution to the conference.**

***Keywords*—photo tagging; face detection; full-body detection; feature extraction; multi-modal search**

I. Introduction

The development in portable devices, which are nowadays equipped with a digital camera led to a need for users to organize their photographs in an effective and efficient manner. There are many public web photo galleries that offer management of photo collections, where some of them provide users with advanced functionality such as automatic people tagging. Examples are Google+ Photos (Picasa) and Apple's iPhoto with the iCloud service where face recognition is used.

In this paper, we propose a framework that goes further and exploits not only face recognition, but also figure (full-body) recognition for automatic management of tags of people. The motivation for such a system is to improve tagging of people who were captured in a posterior position (looking away from camera) so their face cannot be obtained. Assuming the fact that people present at an event usually do not change their clothing during that event, we can take detections of people in the same clothing as the presence of the same person and associate them with a tag saying his or her name, so easing the process of tagging people in photo collections.

The remaining part of this paper is organized as follows. We discuss related work in the next section. In Section III, we describe our proposal and its substantial parts in detail. Evaluation on real-life datasets is given in Section IV. The paper concludes with future directions drawn in Section V.

II. Related Work

The MediAssist system proposed in [1] exploits the idea of using body clothing to improve person identification too. However, they do not detect and recognize human figures in photos but rather extract body patch from the location of person's head. Clothing in the body patch region is used to improve quality of face recognition. By analogy, the authors in [2] define a body region based on the person's face position in an image. An RGB histogram is then obtained from the clothing in body region. Finally, an extrapolation technique to obtain upper-body bounding box is given in [3].

Since the figure detection in still photos is much more challenging than detection in richer sources, e.g., thermal images [4] or video streams [5], we focus on figure detectors in more detail. Many figure or pedestrian detectors exploit Histogram of Oriented Gradients (HOG) features [6]. Follow-up papers improve it by combining HOG with other features, e.g., color channels and histograms of flow [7]. An optimization of HOG to multiscale gradient histograms is introduced in [8], where the scale-space image pyramid is approximated to increase the detection speed. Figure detection based on independent body-part detectors is proposed in [9]. They define a deformable model of parts to create a detector not only for figures but in general for various kinds of objects. This principle is used to segment people in 3D movies [10]. Another approach [11] is based on local binary patterns and its compressed variants. The authors show that this technique outperforms HOG. Figure detection reliability is greatly improved by applying tracking to solve people occlusions effectively [5]. A recent survey [12] of figure detectors gives the reader a complete insight into this problem.

III. Framework for Person Recognition

We propose a generic framework for fusing face and figure feature modalities to significantly improve effectiveness of automatic tagging persons in still images organized in personal photo collections. Figure 1 depicts the proposed framework. The users shown in the figure communicates with the framework by making several requests.

First, the photo-collection-upload request and its processing represent the core of framework. It issues an automatic process of recognition and eventual person tagging. It consists of *detection phase* where faces and figures are localized in the photos, *visual feature extraction phase* where specific descriptors capturing visual appearance in detected regions are obtained, and *clustering phase* where such descriptors are compared by a similarity function to create clusters of the same person. These phases are implemented in independent modules emphasized in blue in the figure.

Second, the result of automatic clustering of faces and figures of the same person may not be perfect, so is not even in Google+ Photos, so requests to manage the tags can be made. It includes naming not-yet-known people, removing false positives in clusters (pictures of different people) and merging separate clusters of the same person.

Third, since the process of person tagging is inherently based on comparing visual features in detected regions, i.e., on similarity, the last request a user can make is a similarity search (image content-based retrieval).

In the following, we focus on the automatic cluster creation and its phases, which form the core of the whole proposed framework.

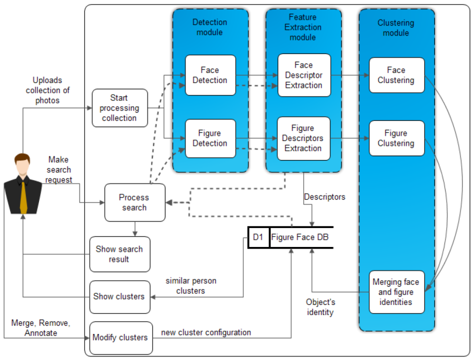


Fig. 1. Schema of framework for people tagging in general photo collections.

A. Detection Module

We use two independent detectors to localize person faces and their figures, which is convenient twofold. It allows running detections in parallel and their implementations can be any of the state-of-the-art methods. The output of the detectors is assumed to be a list of image regions that contain a face or a figure. To detect person faces, we use the Luxand SDK [13] that provides good detection quality, but any other face detector can be used.

For the task of figure detection, we decided to use a method based on Edgelets [14]. This method was designed to work in crowded scenes and an individual human is modeled as an assembly of natural body parts, for each a particular detector is trained by adaptive boosting. This method offers good performance when person is occluded and it is more tolerant to pose and viewpoint change. After training several strong classifiers for detecting individual figures, each image is scanned by a window in scale-space and the classifiers are evaluated. Due to this windowing approach, more detections of the same body can appear, so we join such detections if their overlap exceeds 72%, as defined in (1).

$$\frac{area(R_1 \cap R_2)}{min(area(R_1), area(R_2))} \geq 0.72 \qquad (1)$$

An example of detections and the final merged region is given in Figure 2(b) and Figure 2(c). This merging procedure has a negative effect when large occlusions of bodies appear without all faces being detected, see Figure 2(b). But, this is referred to as a hard problem [7]. To decrease the number of false detections, we filter out all detected regions that cannot be merged with another region. On the other hand, if there is a face detected that coincide with a region of figure detection, it is not filtered out. Both these cases are depicted in Figure 2(e) where both the detections are candidates for filtering out by the first rule. But, the left region is not discarded thanks to its intersection with a face detection, so we take it as reliable enough.

B. Feature Extraction Module

Regions containing persons' faces or their figures detected in uploaded images are passed to the extraction module where descriptors covering visual characteristics are obtained. We use the Luxand SDK again since it offers a high quality face descriptor and a similarity function that perform person identification effectively. This is also confirmed by a comparative study [15] where Luxand SDK in ver. 2.0 exhibited very good values of false rejection and false acceptance rates in a person identification task. Its competitor VeriLook, ver. 4.0 by Neurotechnology, commercial software, offers the same performance but we did not have it licensed. The publicly-available software OpenCV exhibited worse false acceptance rate.

For figure extraction we used the clothing patch covering the central part of body torso. In [14], the torso is defined as the middle part of the whole detected body constrained from top and bottom. To capture the person's clothing as precisely as possible, we have modified this constraint to 0.32–0.58 of the full-body region height and 0.30–0.70 of its width. This was experimentally verified that it maximizes the area of clothing patch while minimizes the influence of background. An example is given in Figure 2(c). A more sophisticated technique to extract clothing based on segmentation can be used [10], [16]. Having obtained a region with clothing patch, we extract one visual descriptor capturing the colors and edges in the clothing patch. In particular, we use a combination of descriptors from the MPEG-7 standard [17]. An experimental evaluation on selecting the best combination is given in Section IV.

Finally, the extraction module produces descriptors consisting of the position and extent of the detected region and the visual descriptor itself for each of the detected faces and human figures. The position and extent are important not only for the clustering module, but also for displaying detections to the user.

C. Clustering Module

This module is responsible for fusing individual detections and their feature descriptors to form groups of images capturing the same person, so a final tag (e.g., person's name) can be assigned to it.

First, all detected faces are separated into clusters by evaluating the Luxand's distance function and the face descriptors whose pair-wise distance is less than 0.14 form a cluster, i.e., describe the same person face. This constant has been experimentally set. In case a different face descriptor or a similarity function is used, this constant must be updated appropriately. Next, the database of known person faces can be searched to identify them and assign their names directly. Currently, we have not implemented such identification yet.

Second, the module proceeds to cluster all detected figures by analogy. For the specific setting of clustering threshold constant on distance and the distance function, please refer to experiments in the next section. Next, the figure clusters are identified by finding correspondences between figure regions and face regions. In particular, we test each figure region in a cluster whether a face region in the same image can be associated with it or not by applying the formula in (2).

$$\frac{area(R_{face} \cap R_{top\_body})}{min(area(R_{face}), area(R_{top\_body}))} \geq 0.10 \qquad (2)$$

It takes the top third of the figure region containing head and shoulders (R_{top_body}) and the face region (R_{face}) and tests their overlap to be at least 10%.

In both the clustering phases, the original image ID from which the detections come, is respected. It obviously assumes that the same person cannot reappear within the same photo, so no two figure nor face detections within the same image can emerge in the same cluster.

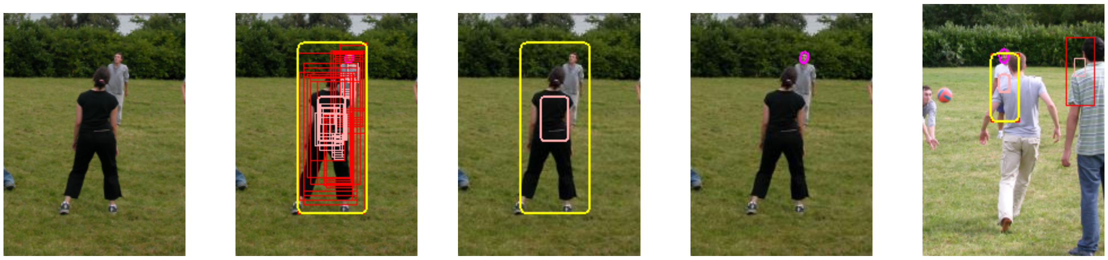


Fig. 2. Figure and face detection result on an image from INRIA person dataset [6]. All detections returned by the Edgelets-based detector are emphasized in red boxes, while the final figure detections after merging are in yellow boxes. The clothing patches (pink boxes) are defined as a region 0.32–0.58 and 0.30–0.70 of height and width of the detection box, respectively. In (a), the original image. In (b), all figure detections with the final detection after merging in yellow; the occluded people are incorrectly merged here. In (c), resulting figure detection with the clothing patch region situated in the middle. In (d), all detected faces. In (e), filtering out figure detections that cannot be merged except detections intersecting a face detection.

IV. Experimental Results

In this section, we give details about training the figure detector and clothing patch extraction since these parts we had to research in order to implement the whole framework proposed in this paper. We also include experience from the clustering people for the task of assigning tags.

A. Figure Detection

To train the figure detector, we used the MIT pedestrian dataset [18] consisting of 914 person images as the positive examples and 1,886 images from the INRIA person dataset [6] as the negative examples.

We tested the quality of trained detector on the ETH person dataset [5], which also includes ground-truth files containing annotations of full-body regions. There are 1,201 person figures in 196 images and our detector correctly identified 817 person figures (68%) and had 29.6% precision (there were other 1,943 detections not containing person figure). We attribute the high number of false positive detections to using a small training set during training detector classifiers. Deeper analysis of this is our future research direction.

B. Clothing Patch Feature Extraction

The task of identifying the person based on their clothing was next challenge. We decided to use a set of global visual descriptors defined in MPEG-7 standard [17]. First, we defined the clothing patch region (see Figure 2(c)) based on our experience with the detector. Second, we picked color structure, scalable color, color layout and edge histogram, since they work on small images and capture not only color and also other visual features. We tested them on the ETH person dataset. From various trials ranging from individual descriptors to their weighted combinations and following the paper [19], we concluded with the combination of scalable color, color structure and edge histogram normalized and weighted in the ratio 5:2.5:1, respectively. This combination reached 0.797 value of Mean Average Precision (MAP), see Figure 3. For space constraints, we do not include other results. The distance constant we used to cluster figures clothed similarly, was set to 1.28.

C. People Tagging

We tested the quality of tagging people on a subset of ETH dataset. We selected 477 images taken from the BAHNHOF sequence. The values of distance used to cluster face and figures were set quite strictly, so the automatic process created 290 clusters of faces and figures. There were on average 7.8 clusters per person, a value which was obtained by manual checking the resulting clustering. On the other hand, eight clusters (out of 290) contained different persons, so the clustering failed here. It was caused mainly by people in black coats and their occlusions with tree trunks. The implemented system then allows the user to manage the clusters manually, i.e., to merge clusters of the same person, to name the person, etc.

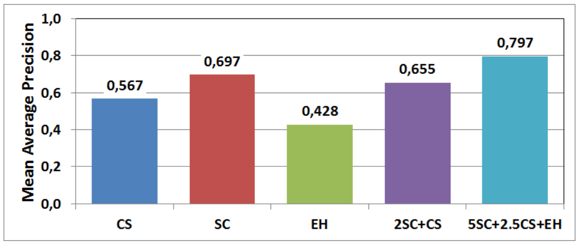
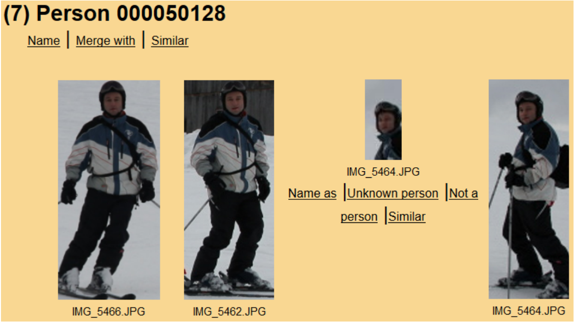


Fig. 3. Comparison in MAP of different global descriptors and their combinations. The last column depicts variant used by ourselves.

We have also tested the prototype implementation on a small personal holiday collection that contained 24 photos of people indulging winter activities. Original photographs were of 18 megapixels but we had had to down-sample them to around 400 by 300 pixels since our figure detector was trained on small-resolution images, as have been mentioned above. This toy dataset contains 76 figures, each at least 220 pixels tall (in original resolution), and 54 faces. Many faces are covered with skiing goggles, which makes it very challenging for common face detectors. In these photographs, 26 distinct people were shot, 15 of which only once.

Our face detector revealed 7 faces only, where all were true positives. The figure detector correctly bounded 54 figures out of 78 detections. They were grouped in 5 correct clusters, thus they contained photos of the same person. Three clusters contained different people, but they were all wearing very similar clothes – red, black or red/black jackets. One cluster consists of a face-figure pair of one person. Next, seven clusters can be considered as mixtures of false-positive and true-positive detections, where the clothing patches are alike. Finally, the other detections were not grouped at all, so they resulted in “one-detection” clusters. An example of an automatically created cluster is given in Figure 4. The complete results are



(7) Person 000050128

Name | Merge with | Similar
IMG_5464.JPG
Name as | Unknown person | Not a person | Similar
IMG_5466.JPG
IMG_5462.JPG
IMG_5464.JPG

Fig. 4. A cluster of grouped images of the same person. The person's tag has been generated from database cluster ID.

available at http://disa.fi.muni.cz/mmedia2014/.

We tested Google+ Photo by creating an album out of the original high-resolution photos. Google's software detected 16 faces and clustered them into 12 clusters. By manual verification, these detections correspond to 7 people. We attribute these results to the Google's policy to provide its users with high-precision face detections. Surprisingly, no faces were detected in the down-sampled photos.

V. Conclusions

We proposed a framework that combines a face and figure (full-body) detector to recognize people with the aim of providing people tagging in user photo collections. The contribution of this paper is in testing various descriptors for comparing clothing patches to recognize similar clothing, which in other words, leads to identification of the same person in different photographs. This, of course, requires the assumption that people do not change their clothes within a short period of time in which a social event takes place. The other and main contribution is in implementing a prototype using state-of-the-art techniques for face and figure detections and fusing these two modalities into one system. The prototype is available at http://disa.fi.muni.cz/mmedia2014/.

This preliminary prototype can be improved fourfold. First, the face detection module can be changed to use a multi-view face detector [20], which was successfully used in a recent paper on finding actors in movies and assigning their name and actions from movie scripts [3]. Second, preparing an Edgelets detector for not only full-body detections, but also an upper-body detector on a bigger training data set is our next goal. Third, the personal holiday photos do not contain overcrowded scenes very often, so a better alternative would be to apply a detector based on histogram-of-gradient features and latent support vector machines [8]-[9]. Fourth, person occlusion (see Figure 2(b)) can be partly eliminated by training a separate head detector to avoid merging two figure detection if they contain two different heads.

Finally, the proposed system can be used to assign person tags very easily having a better figure detector with low rate of false positives.

Acknowledgment

Vlastislav Dohnal was supported by the Czech Science Foundation, project no. GBP103/12/G084. Alexander Matecny was supported by the Ministry of the Interior, Czech Republic, project no. VG20122015073.

References

[1] N. O'Hare and A. Smeaton, "Context-aware person identification in personal photo collections," Multimedia, IEEE Transactions on, vol. 11, no. 2, 2009, pp. 220–228.

[2] L. L. Presti, M. Morana, and M. L. Cascia, "A data association algorithm for people re-identification in photo sequences," Multimedia, International Symposium on, 2010, pp. 318–323.

[3] P. Bojanovski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic, "Finding actors and actions in movies," in IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 3-6, 2013. IEEE, 2013, pp. 1–8.

[4] M. Correa, G. Hermosilla, R. Verschae, and J. Ruiz-del Solar, "Human detection and identification by robots using thermal and visual information in domestic environments," Journal of Intelligent & Robotic Systems, vol. 66, no. 1-2, 2012, pp. 223–243.

[5] A. Ess, B. Leibe, K. Schindler, , and L. van Gool, "A mobile vision system for robust multi-person tracking," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Press, June 2008, pp. 1–8. [Online]. Available: http://www.vision.ee.ethz.ch/~aess/dataset/[retrieved:Dec.,2013]

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in International Conference on Computer Vision & Pattern Recognition, vol. 2, June 2005, pp. 886–893. [Online]. Available: http://pascal.inrialpes.fr/data/human/[retrieved:Dec.,2013]

[7] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 1030–1037.

[8] P. Dollar, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in Proceedings of the British Machine Vision Conference (BMVC). BMVA Press, 2010, pp. 68.1–68.11, doi:10.5244/C.24.68.

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, no. 9, 2010, pp. 1627–1645.

[10] K. Alahari, G. Seguin, J. Sivic, and I. Laptev, "Pose estimation and segmentation of people in 3d movies," in Proc. IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1–8.

[11] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, "Discriminative local binary patterns for human detection in personal album," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[12] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), vol. 34, no. 4, 2012, pp. 743–761.

[13] Luxand, Inc., "Luxand face SDK 4.0," 2005-2013. [Online]. Available: http://www.luxand.com/facesdk/[retrieved:Dec.,2013]

[14] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in IEEE International Conference on Computer Vision, ICCV 2005, vol. 1, 2005, pp. 90–97.

[15] N. Degtyarev and O. Seredin, "Comparative testing of face detection algorithms," in Image and Signal Processing, ser. LNCS. Springer Berlin Heidelberg, 2010, vol. 6134, pp. 200–209.

[16] A. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[17] B. Manjunath, P. Salembier, and T. Sikora, Eds., Introduction to MPEG-7: Multimedia Content Description Interface. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[18] C. Papageorgiou, T. Evgeniou, and T. Poggio, "A trainable pedestrian detection system," in Proceeding of Intelligent Vehicles, October 1998, pp. 241–246. [Online]. Available: http://cbcl.mit.edu/software-datasets/PedestrianData.html[retrieved:Dec.,2013]

[19] M. Batko, P. Budkov, and D. Novk, "Cophir image collection under the microscope," in Proceedings of the 2009 Second International Workshop on Similarity Search and Applications. Washington, DC, USA: IEEE Computer Society, 2009, pp. 47–54.

[20] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2879–2886.