Human Relative Position Detection Based on Mutual Occlusion

Víctor Borjas, Michal Drozdal, Petia Radeva, and Jordi Vitrià

Universitat de Barcelona,
Facultat de Matemàtiques & Centre de Visió per Computador,
Campus UAB

Abstract. In this paper, we propose, within the field of automatic social context analysis, a novel method to identify the mutual position between two persons in images. Based on the idea that mutual information of head position, body visibility and bodies' contour shapes may lead to a good estimation of mutual position between people, a predictor is constructed to classify the relative position between both subjects. We advocate the use of superpixels as the basic unit of the human analysis framework. We construct a Support Vector Machine classifier on the feature vector for each image. The results show that this combination of features, provides a significantly low error rate with low variance in our database of 366 images.

1 Introduction

When two persons are facing an audience or facing a camera for a shot, it is common that one of the subjects occludes part of the other's body. This occlusion may be originated by both physical or social attributes. For example, physically shorter people (usually, females or children) tend to stand in front of the other subject producing the occlusion, or socially more important people tend to stand in a closer plane to the camera or audience. Fig. 1 sketches the situation addressed. In these images we can observe couples in a frontal view from camera. In these cases, the relative position can denote political, hierarchical or leadership status between them. These social relations can be of great importance when analyzing social and cultural behavior like social hierarchy, political importance or leadership. Multiple characteristics of the scene are involved like head and body position, clothing color, pose analysis, and many others.

Up to our knowledge, the amount of papers related to analyze faces as a group instead of individuals, and the physical mutual relation of people within a scene is surprisingly low, usually in the context of identifying people in collection of photos like in [5], [7] and [8]. In [6], the authors analyze people in images using contextual features extracted from relative head position and size that encapsulate the structure of the scene to perform tasks like demographic recognition, calculating scene and camera parameters, and event recognition. They show that factors like age or gender influence the position of subjects in an image.

Fig. 1. (a) and (d) Original images; (b) and (e) Head Detection boundaries in blue, intersection zone in red and Body Segmentation labels: white for Background, black for Left and grey for Right; (c) and (f) Curve Fitting in the intersection zone

Our goal is to develop an automatic method to detect relative position of people in frontal images in terms of mutual human body occlusions. First, using a Head Detection Algorithm [9], faces are detected (see Fig. 1 (b) and (e)). Next, the image is then converted into superpixels to act as the basic element of all operations. Skin superpixels are identified and eliminated from the image to avoid confusion and misclassification. The parameters of three RGB Gaussian Mixture Models (GMM) corresponding to the background and both persons are computed. A previously stated body hypothesis is projected on the image to delimit body prior boundaries. An energy minimization cut is performed in an undirected graph, where each node is a superpixel (Fig. 1 (b) and (e)). Edges are extracted from the "uncertainty" zone between both bodies and curves are fitted to analyze the image structure (Fig. 1 (c) and (f)). Finally, a Support Vector Machine (SVM) is trained to classify the mutual position.

2 Our Approach

Given the problem of detecting mutual position of persons, we propose to extract a feature vector that represents scene in terms of head position, body visibility and bodies' contour shapes.

2.1 Superpixels

In order to simplify the image into a new representation that is more meaningful and easier to analyze, we use the geometric flows method [2] to extract the superpixels from the images. Superpixels represent regions of pixels that have nearly the same color and brightness. This algorithm produces segments that on one hand respect local image boundaries, while on the other hand limit under-segmentation through a compactness constraint. Pixels are guaranteed to maintain an uniform size along the image, giving an advantage for GMM calculation and skin detection. Fig. 2 (a) and (b) show an example of this extraction.

2.2 Skin Detection

Images with high percentage of skin may lead to misclassification since they reduce the effectiveness of the GMM as a basic superpixel classifier. To this

Fig. 2. (a) Original image. (b) Superpixels. (c) Skin detection. (d) Connected Superpixel Graph. (e) Closer Visualization of Graph

purpose, we disable skin pixels in the images for contributing to the classification process. According to [10], one of the fastest and more reliable ways of identifying skin pixels in an image is presented in [3]. Three basic threshold operations (see Eq. (1)) over three terms built from RGB colors proved to have above 90% of precision. An example of the resulting skin class mask is shown in Figure 2 (c).

$$\frac{3br^2}{(r+g+b)^3} > 0.1276 \ \& \ \frac{r+g+b}{3r} \frac{r-g}{(r+g+b)} \leq 0.9498 \ \& \ \frac{rb+g^2}{gb} \leq 2.7775 \quad (1)$$

2.3 Body Visibility

Aiming to extract an attribute that defines body visibility, we need to perform body segmentation. Using GMM to represent section or classes in an image have been widely used and proved to be effective. In [11], the authors perform background/foreground segmentation based on iterative graph cuts via energy minimization. He aims for a quick foreground segmentation simplifying user interaction. In respect to that, a foreground and background GMM is defined from user input and a k value (GMM correspondence) to every pixel is assigned. Secondly, GMM means and covariances are updated, with the corresponding weights. To perform final segmentation, the author performs graph cut via energy minimization as in [13] giving each pixel in the image a label l . This process is repeated iteratively until the minimum energy is reached.

In our algorithm, a weighted GMM model in RGB color space is used as initial pixel classification. Models are built for *Background*, *Left Person* and *Right Person*. We follow a practice that is already used for soft segmentation [14]. Each GMM is taken to be a full-covariance Gaussian mixture with K components (we use $K = 2$ for foreground/bodies and $K = 3$ for background).

GMM 'A Priori' Classification. In previous recognition work using clothing [4], either a rectangular region below the face is assumed to be clothing, or the clothing region is modeled using operator-labeled clothing from many images. Similar to [1], we use a non-rectangular mask (Fig. 3(a)), computed from a set of body labeled images, projecting it into the image scaling it according to each face size. From this projection, we create two masks identifying both body hypothesis useful for further operations as shown in Fig. 3 (c) and (d).

Fig. 3. (a) Body Mask. (b) Turbopixel segmentation. (c) and (d) Left and Right body mask projections. (e) Background mask. (f) Foreground mask. (g) and (h) Left and Right weighted masks for GMM training.

The hypothetical foreground and background regions in the image are obtained joining both body hypothesis as shown in Fig. 3 (c) and (d). Skin super-pixels are added to the background as explained in Section 2.2. Both foreground GMM (Left and Right) are then computed assigning weights to the pixels, assuming that the most-left pixels within the foreground belong to the left-hand person. The pixel weights decrease as the position in the horizontal axis increases for the Left GMM and viceversa. Fig. 3 *Bottom Row* shows an example of how weights behave within the foreground.

Finally, GMM for each label are trained using the corresponding weights and pixels stated previously. The weighted means μ_l and weighted covariance matrices Σ_l for each model are stored for further likelihood calculation. Assuming weights are normalized $\sum_{i=1}^N \omega_i = 1$, where N is the number of pixels, means and covariance matrices for each label l are computed as follows:

$$\mu^l = \sum_{i=1}^N \omega_i^l x_i, \quad \{\Sigma_l\}_{jk} = \frac{\sum_{i=1}^N \omega_i^l}{(\sum_{i=1}^N \omega_i^l)^2 - \sum_{i=1}^N (\omega_i^l)^2} \sum_{i=1}^N \omega_i^l (x_{i,j}^l - \mu_{j,k}^l)(x_{i,k}^l - \mu_{j,k}^l) \quad (2)$$

where x is the RGB value of each pixel and l denotes the label (Background, Left or Right).

Body Segmentation and Graph Cut via Energy Minimization. An energy function E is defined so that its minimum should correspond to a good segmentation guided by the observed foreground and background GMM, and reflecting solidity of objects. This is captured by a "Gibbs" energy of the form: $E(L, \theta, Z) = U(L, \theta, Z) + V(L, Z)$, where $L = \{l_i\}_{i=1, \dots, N} = \{1, 2, 3\}$ is the set of possible labels, θ is the set of GMM and Z is the set of pixels. The data term U evaluates the probability of assigning label l_i to pixels in Z given the GMM θ : $U(L, \theta, Z) = \sum_{z_i \in Z} -\log(P(z_i | l_i))$. The smoothness term can be written

as: $V(L, Z) = \sum_{(m,n) \in H} [l_m \neq l_n] \exp(-\beta(z_m - z_n)^2)$ where H is the set of connected nodes, $[\cdot]$ acts as a factor that allows summation only pairs of nodes with different assigned label l and β is a smoothness term also known as Ising prior. This energy term encourages coherence in regions of similar RGB values. Good results are obtained by defining nodes to be connected if their distance in image coordinates is below a given threshold t :

$$t = \frac{\sum_{i=1}^N d(z_i)}{N} \quad (3)$$

where $d(\cdot)$ is a function that returns the longest possible diameter of a given superpixel. This limits the neighbor to those nodes within a diameter distance in any direction as sketched on Fig. 2 (d) and (e).

A reliable value of β is $\beta = (2\langle(z_m - z_n)^2\rangle)^{-1}$ where $\langle z_m - z_n \rangle^2$ represents the Euclidean distance in the RGB color space between pixels z_m and z_n [13]. Minimization of $E(L, \theta, Z)$ is done using the standard minimum cut algorithm defined in [13].

2.4 Curve Fitting

Once the bodies are segmented in the image, we perform a curve fitting process on important edges along the intersection of them. Assuming an overlapping between body hypothesis, we use their intersection as the "uncertainty zone" (see Fig. 1 (b) and (e)). We extract this zone from the labeled image and detect edges using Sobel filters. Considering that edges belonging to background should not be taken into account for curve fitting, we filter them by performing basic morphologic operations in the edge image (See Fig. 1 (c) and (f)). This procedure leaves only edges that correspond to boundaries between both bodies.

Afterwards, a polynomial curve is fitted on the edge points. In this case, we used second order polynomial since we are aiming for a simple quadratic function curve that corresponds to shoulder-like shapes. Once the curve is fitted, mean normal vectors are computed. Since the direction and magnitude of these vectors represent the shape or type of curve, this vector is directly used as a feature, n^* .

2.5 SVM Training

We use a SVM classifier with a third grade polynomial Kernel function to classify each new image as *Left* or *Right* human occlusion. Each image is then represented as a 3D feature vector:

$$F = [y^*, r^*, n^*] , \quad y^* = \frac{y_i - y_j}{\frac{a_i + a_j}{2}} , \quad r^* = \frac{\sum_{\{z_i | l_i = right\}} 1 - \sum_{\{z_j | l_j = left\}} 1}{N} \quad (4)$$

where y^* is the normalized difference in vertical position of both detected heads in the image. Normalization is done over the average scale $\bar{s}$ of heads obtained as

the head area $a_i = w_i \cdot h_i$, where w_i represents the head width and h_i represents the head height in the image. The term r^* represents the difference in percentage of labels in the "uncertainty zone" graph cut segmentation. Normals n^* come directly from the edge detection and curvature detection step in Section 2.4. Figure 4 sketches the three components of the feature vector.

Fig. 4. (a) Relative height difference (y^*), (b) Body visibility likelihood (r^*), and (c) Contour shape estimated from the curve fitting (n^*)

3 Experiments

3.1 Database

We constructed a dataset that consists of 366 images with ground truth annotations in the form of *Left* and *Right* overlapping. Each image includes only two persons in frontal view to the camera and one is partially occluding the other. Since there is no proper public domain database, we created one by randomly downloading from internet and cropping the images in situations when more than two persons appeared close to each other and an overlapping is possible to occur. The database is available for downloading in the first author's website.

3.2 Evaluation

We performed a ten-fold cross-validation measuring the mean score and variance of all possible combinations of features. Fig. 5(a) shows how the classifier performs, noticing that the highest mean score and the lowest variance, are obtained by combinations of the explained features, y^* , r^* and n^* . Additionally, very low variance values are important to notice.

Fig. 5(b) sketches the precision of our system in function of number of superpixels in which the image is represented. It shows that the best results are obtained with 200 superpixels per image, corresponding to 10-15 superpixels per face. This is an important fact because it tells how general a representation of an image can be avoiding image details and still be able to perform GMM, graph cut segmentation and curve fitting.

Failures can occur mainly in three cases shown in the *Right Column* of Fig. 6: (a) when skin detection is not precise, either it gives false positives like labeling clothing as skin, or false negatives like not labeling a face as skin; (b) both persons clothing has the same distribution of color, like both dressed in black which can lead to a weak segmentation in "uncertainty zone"; (c) uncommon human poses, lead to a wrongly determined "uncertainty zone", therefore producing a wrong segmentation and an inaccurate curve fitting.

Fig. 5. (a) Mean Precision and Mean Precision Covariance. (b) Mean Precision as function of number of superpixels.

Fig. 6. From Left to Right: Original image with Head Detection; Superpixel representation; Graph Cut result in greyscale labels and "Uncertainty Zone" in red; and Curve Fitting in blue. (a), (b) & (c) sketch correct classification. (d) Too much skin in the image leads to lack of information for further processes. (e) Color similarity between both persons' clothes produces graph cut errors. (f) Uncommon human pose leads to curve fitting errors.

4 Conclusions and Future Work

In this paper, up to our knowledge, for the first time an automatic detection of mutual occlusion of persons is addressed. We presented a novel method to identify the mutual position between two persons in the same plane of an image. Our system relies on an efficient feature vector construction from both simple and complex characteristics of the scene. It resulted in a reliable classification system in uncontrolled datasets obtaining an accuracy of 88%.

The framework described here allows for exploration in a wide set of directions to enrich the model like: new GMM class for skin recognition and interpretation on possible arm and head poses, learning the likelihoods of edge orientation according to face or upper body appearance.

Acknowledgments. This work was partially supported by MICINN grants TIN2009-14404-C02, CONSOLIDER-INGENIO 2010 (CSD2007-00018) and the AGAUR grant to SGR696.

References

1. Sivic, J., Zitnick, C., Szeliski, R.: Finding people in repeated shots of the same scene. In: Proc. BMVC (2006)
2. Levinshtein, A., Stere, A., Kutulakos, K.N., Fleet, D.J., Dickinson, S.J., Siddiqi, K.: TurboPixels: Fast Superpixels Using Geometric Flows. TPAMI (2009)
3. Gomez, G., Morales, E.F.: Automatic feature construction and a simple rule induction algorithm for skin detection. In: Proc. ICML Workshop on Machine Learning in Computer Vision
4. Gallagher, A., Chen, T.: Clothing cosegmentation for recognizing people. In: Proc. CVPR (2008)
5. Gallagher, A., Chen, T.: Using group prior to identify people in consumer images. In: Proc. CVPR SLAM Workshop (2007)
6. Gallagher, A., Chen, T.: Understanding Images of Groups of People. In: Proc. CVPR (2009)
7. Naaman, M., Yeh, R., Garcia-Molina, H., Paepcke, A.: Leveraging context to resolve identity in photo albums. In: Proc. JCDL (2005)
8. Stone, Z., Zickler, T., Darrell, T.: Autotagging facebook: Social network context improves photo annotation. In: Proc. CVPR, Internet Vision Workshop (2008)
9. Drozdal, M., Hernández, A., Segu, S., Baro, X., Escalera, S., Lapedriza, A., Masip, D., Radeva, P., Vitria, J.: Combining Detectors for Human Layout Analysis. In: Proc. ECCV, The Pascal Visual Object Classes Challenge Workshop (2010)
10. Vezhnevets, V., Sazonov, V., Andreeva, A.: A Survey on Pixel-Based Skin Color Detection Techniques. In: Proc. GRAPHICON (2003)
11. Rother, C., Kolmogorov, V., Blake, A.: GrabCut, interactive foreground extraction using iterated graph cuts. ACM Trans. SIGGRAPH (2004)
12. Fulkerson, B., Vedaldi, A., Soatto, S.: Class segmentation and object localization with superpixel neighborhoods. In: Proc. ICCV (2009)
13. Boykov, Y., Jolly, M.: Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In: Proc. ICCV (2001)
14. Ruzon, M., Tomasi, C.: Alpha estimation in natural images. In: Proc. CVPR (2000)