Human Body Part Classification in Monocular Soccer Images

Andreas Bigontina, Michael Herrmann, Martin Hoernig and Bernd Radig
Image Understanding and Knowledge-Based Systems
Technische Universität München
Boltzmannstr. 3, 85748 Garching, Germany
Email: {bigontia,herrmmic,hoernig,radig}@in.tum.de

***Abstract*—This paper addresses the problem of finding body parts in images, which can be an essential first step for body pose estimation. The core component of the presented method is the pixel-based classification of body parts using Random Forests. This technique is applied to find the body part positions of soccer players in broadcast images. As this approach is usually used with depth data, we analyze how this method can be adapted to work with monocular images. In this context we identify the image representations with the best classification results. Although monocular images leave some ambiguities, our approach to body part classification achieves satisfying results: 90.32% of the pixels in the test set are correctly classified.**

I. Introduction

The aim of articulated pose estimation is to determine the position of body parts and/or joints of a model of the human body in 2D or even 3D space. Hence, finding the position of body parts in an image, which is in the focus of this paper, can be seen as an essential part of a pose estimation algorithm. Combined with a skeleton fitting method it can be used to estimate 2D or 3D poses (e.g. by applying an inverse kinematic technique similar to Wei et al. [1]). In this paper we examine the particular application of detecting the body parts of soccer players in monocular images of soccer matches. As we work with cropped images of players from distance shots, we require a solution that can handle low resolution images.

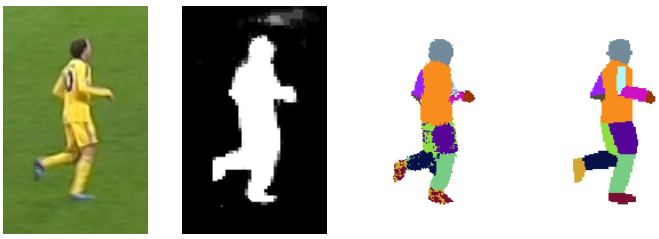
Despite its use as part of a pose estimation system, the pixel-based classification yields precious information on its own. During the analysis of a soccer video sequence, for instance, it might help to indicate whether the ball was hit by the player's foot, head, or hand. But the presented method can also be applied in other areas, given a monocular image and a person's silhouette. Some of the many areas of application are autonomous cars, human-robot interaction, surveillance, video indexing, and sports. Especially, when already employed cameras should be used or already recorded material should be analyzed, a solution that works with monocular images is required.

In summary, given the surrounding bounding box of a player in an image, our method performs two successive steps: In the first step we estimate the body orientation of the player. The result serves as input for the second step: For each pixel within the bounding box we determine the body part of the player to which the pixel most probably belongs.

Our proposed approach builds upon the work of Shotton et al. [2] who developed the pose estimation system for the Kinect gaming platform [3]. Similar to their approach, a Random Forest is used to classify each pixel of an image based on the state of the pixels in the neighborhood. The aim is to assign a body part or background class to every pixel in the image, as depicted in Fig. 1. This paper focuses on finding alternative image features to replace the missing depth information that was originally provided by the Kinect sensor. Therefore, several preprocessing steps are examined.

In addition, the body part classification of soccer players is more challenging than the classification of Kinect players since there is a greater variation in viewpoints. Therefore, an idea of Andriluka et al. [4] is followed who suggest to estimate the viewpoint beforehand. For this task, the orientation of a player is categorized into eight classes. A Random Forest is trained to estimate the orientation class of a player using the same techniques as for body part classification.

The main contribution of this paper is the adaption of the algorithm by Shotton et al. [2] to work with monocular images and the comparison of image features for this task. Furthermore, the problem of finding the orientation of a player is examined and corresponding results are presented.



(a) (b) (c) (d)

Fig. 1: Body part classification: (a) original player image, (b) silhouette retrieved using [5], (c) estimated body parts, (d) ground truth annotations

A. Related Work

The problem of finding the pose of a person from sensor data has been investigated by many researchers before. For surveys on this topic see the work of Moeslund et al. [6] or Poppe [7]. With the upcoming of the Kinect gaming platform [3], developed by Microsoft, an accurate and fast pose estimation algorithm was presented [2] that is working on depth information. The main advantage of their algorithm

is a fast classification using a Random Forest with the idea of classifying every pixel separately. This allows for parallelization and implementation on GPUs, as shown by Sharp [8]. The presented approach differs from Shotton et al. [2] as it does not rely on depth information but works with monocular images.

Nevertheless, there are also several solutions for monocular images presented in literature. Many of these rely on the pictorial structures model, originally introduced by Fischler and Elschlager [9] and later applied to human pose estimation by Felzenszwalb and Huttenlocher [10]. The final results of a system using their techniques, for instance that of Andriluka et al. [11] who even estimate a 3D pose, are quite impressive. The complex series of actions that is required, however, also comes with a computational burden. For instance, Andriluka et al. [11] report an overall runtime of about 50 seconds for a 167×251 pixel image. In comparison, the method by Shotton et al. [2] can perform in real-time. Wei et al. [1] extended their approach and estimated a pose using the pixel-based classification, still keeping real-time performance. The presented approach clearly differs from classical pose estimation algorithms for monocular images due to the pixel-based classification concept and, thus, has computational advantages.

II. Body Part Classification

As mentioned above, the classification of body parts is a core component of our method. Since images from broadcast videos might have a very low resolution, a rather coarse division into parts is necessary. Furthermore, the body model should be divided into more or less rigid parts to minimize the variation in appearance of each part. We use the following 14 classes: head, torso, left/right lower/upper arm, left/right lower/upper leg, hands, and feet (similar to [12]). The presented method largely follows the idea of Shotton et al. [2]. A Random Forest is used to classify every pixel of an image. Shotton et al. compare depth values of pixels in the neighborhood to decide about the class of a pixel. As such depth information is not available in monocular scenarios the following alternatives are examined: silhouette information, skin, color, gradient images, Haar-like features on gradients, HOG descriptors, pixel positions, and estimated player orientations. Since some of these features might give partial information about certain body parts only, feature combinations are examined as well. The silhouettes are retrieved using the method of Hoernig et al. [5] which is especially applicable in soccer and similar field sports. Silhouettes are very expressive but naturally leave some details unrecognized. Skin features are extracted from the image using the method described by Gomez and Morales [13] and are intended to complement the silhouette information. Additionally, raw color values are examined as well to see if the applied Random Forest is powerful enough to extract the most useful information by itself. Another way of finding hidden details inside silhouettes are gradients and their magnitudes. However, the most useful, strong gradients usually exist at thin edges and are therefore hard to utilize by the presented approach. The idea of summing up the gradients within an image area is implemented using Haar-like features (similar to [14]). Finally, Histograms of Oriented Gradients (HOG) [15] are examined, as they have shown good results in human detection.

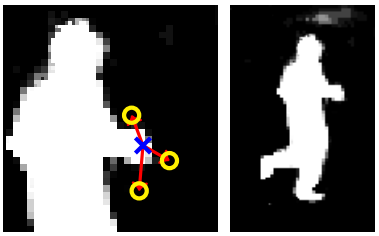


Fig. 2: Illustration of the classification method for silhouettes: the blue cross marks the pixel that has to be classified, the yellow circles mark positions in the neighborhood at which foreground probabilities are compared to a threshold

We trained Random Forests using these features and feature combinations to estimate the orientation of a player. Each node in the forest examines a feature value at a certain position in the image and compares it to a threshold. The most expressive position that contains the most distinctive feature values and the threshold were chosen during the training phase. The estimated orientation can be important for further pose recovery tasks, but it also gives hints about body part classes. Therefore, the result of the orientation classification serves as a feature for the pixel-based body part classification.

While orientation classification is concerned with the orientation of a whole player image, body part classification tries to find a class for a single pixel. Hence, an additional information that is available is the pixel position relative to the player's bounding box. Therefore, it is considered as a feature for the body part classification as well.

For all features, a simple axis-aligned weak learner is used. This means that a simple comparison between a feature attribute value and a threshold is performed in each node of a Random Tree. In the training phase an expressive feature attribute has to be chosen and an appropriate threshold has to be found. For a detailed description of Random Forests see, for example, the extensive work by Criminisi et al. [16].

We trained different Random Forests on features and combinations of these and performed an evaluation using a challenging data set containing 70 images of players in various poses, resulting in millions of test pixels. Surprisingly, a rather simple combination of features showed to perform best: silhouette, color, orientation, and pixel position. Therefore, these four features are described in detail.

A. Silhouette

The silhouette information retrieved using [5] is not binary but provides foreground probabilities for every pixel. The applied weak learner of the Random Forest compares foreground probabilities in the neighborhood of a target pixel to a threshold. Which pixel in the neighborhood should be examined is specified by an offset that has to be found in the training phase. This is illustrated in Fig. 2.

B. Color

Color contains full information, hence, it is more challenging to extract the most relevant information. Similar to the idea of Shotton et al. [2] two different strategies are encoded in

TABLE I: Orientation classification accuracy

Features	Accuracy
Silhouette	51.4%
Silhouette + Skin	49.3%
Silhouette + HOG	40.7%
Silhouette + Gradients	38.5%
Silhouette + Color	35.0%
Color	29.3%
Silhouette + Haar-like	28.6%

TABLE II: Body part classification accuracy

Features	Accuracy
Sil. + Orientation + Position + Color	90.32%
Sil. + Orientation + Position + Gradients	89.54%
Sil. + Orientation + Position + Skin	89.46%
Sil. + Orientation + Position + HOG	89.43%
Sil. + Orientation + Position + Haar-like	89.37%
Sil. + Orientation + Position	89.34%
Silhouette	87.13%
Silhouette + Orientation	87.03%
Color	85.23%

the applied weak learner of the Random Forests. One option is to compare the value of a color channel to a threshold. The other option is to compare the value difference between two pixels regarding one color channel to a threshold. This feature involves choosing one of these options and finding one or two offsets, a color channel, and a threshold during the training.

C. Pixel Position

This paper is only concerned with finding the body parts of players, thus, it is assumed that players are already detected and that their bounding boxes are known. A further improvement of the classification accuracy can be achieved by adding the position of a pixel relative to a players bounding box as a feature (see Table II). To optimally utilize the position data, the Random Forest is provided with redundant information. In particular, a pixel's distance to the top, left, right, and bottom of the bounding box are provided, as well as the coordinates relative to the center. The applied weak learner selects one of these measurements and compares it to a threshold.

D. Player Orientation

As mentioned above, the orientation of a person is estimated as well. The Random Forest for body part classification contains a weak learner that compares the estimated probability of one of the eight orientation classes to a threshold.

III. Experiments

To evaluate the performance of the proposed system and to find the best feature or feature combination an extensive evaluation was performed. Concentrating on the special application of broadcast videos of soccer matches a corresponding data set was created that contains 200 images of soccer players from various viewpoints, in various qualities and containing motion blur and similar challenges. The average size of these images is about 78×120 pixels, but every image was scaled to a height of 200 pixels (with fixed aspect ratio) to cope with different image sizes to some extend. The data set was split into 130 training and 70 test images. To enhance the data set a flipped version of every image was added, while assuring that players from test and training set are not mixed up. In all images of the data set body parts were manually annotated on pixel-level.

Concerning the classification of a player's orientation, a Random Forest that uses the silhouette only, achieves the best results for the soccer data set (see Table I). Classifying the orientation is a difficult task, only 51.4% of the player images are classified correctly, however, it is sufficient to support body part classification. Furthermore, it is a useful information for upcoming pose estimation tasks. It is likely that performance could be improved by increasing the number of training samples. This might also help to utilize more complex features better. Although, 130 images yield millions of training pixels for the body part classification, they only yield 130 samples to train the orientation classifier.

As already mentioned above, our most accurate feature combination for body part classification is silhouette, color, position, and orientation. An example result is visualized in Fig. 1. Using this feature combination 90.32% of the pixels in the test set are correctly classified. Here, a pixel is considered correctly classified if the class with the highest probability equals the ground truth class. As can be seen from Table II, the results for other combinations are only slightly worse. Note, however, that a significant part of the pixels of a player image are background and are already indicated as such by the silhouette. In fact, if a simple threshold is applied to the silhouette probabilities and the resulting foreground pixels are classified as torso (the most frequent class), already 79.95% of the pixels in the test set are correctly classified. Thus, a good classifier is expected to improve significantly over this baseline.

The best performing Random Forest within our test set contained 20 trees, which are trained to a depth of 20 but are also limited by a minimal amount of samples per leaf of 92. Every tree is trained on a random subset of the training data. During training only a random subset of the theoretically possible splits is examined as it is common practice for Random Forests. The neighborhood of a pixel that is examined is limited to a 50×50 pixels patch for images of height 200 pixels. An examination of this Random Forest shows that most information is retrieved from the silhouette, while the orientation, the color and the pixel position only improve results slightly.

The classification of orientation and body parts can be performed in less than a second on a recent CPU. As already mentioned, Random Forests can be implemented on GPUs which promises a significant performance improvement.

IV. Conclusion

The performed experiments have shown that it is indeed possible to use pixel-based classification on image data from monocular cameras to estimate the position of body parts with high accuracy, even for low resolution and low quality images. Furthermore, it has been shown that simple features are usually sufficient to receive good results. However, some challenges remain such as distinguishing the left and right body parts, and scoping with erroneous silhouettes, which contain field lines, advertisement banners, or similar background objects.

REFERENCES

[1] X. Wei, P. Zhang, and J. Chai, "Accurate realtime full-body motion capture using a single depth camera," *ACM Transactions on Graphics*, vol. 31, no. 6, p. 1, Nov. 2012.

[2] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Computer Vision and Pattern Recognition*, vol. 2, no. 3. IEEE, 2011, pp. 1297–1304.

[3] Microsoft Corp. Redmond WA, "Kinect for Xbox 360," Redmond WA, 2010.

[4] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in *Computer Vision and Pattern Recognition*. IEEE, Jun. 2010, pp. 623–630.

[5] M. Hoernig, M. Herrmann, and B. Radig, "Real Time Soccer Field Analysis from Monocular TV Video Data," in *11th International Conference on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-11-2013)*, vol. II, Samara, 2013, pp. 567–570.

[6] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90–126, Nov. 2006.

[7] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, Jun. 2010.

[8] T. Sharp, "Implementing Decision Trees and Forests on a GPU," *Computer Vision - ECCV 2008*, vol. 5305, pp. 595–608, 2008.

[9] M. A. Fischler and R. A. Elschlager, "The Representation and Matching of Pictorial Structures," *IEEE Transactions on Computers*, vol. C-22, no. 1, pp. 67–92, Jan. 1973.

[10] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial Structures for Object Recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[11] M. Andriluka, S. Roth, and B. Schiele, "Discriminative Appearance Models for Pictorial Structures," *International Journal of Computer Vision*, vol. 99, no. 3, pp. 259–280, Oct. 2011.

[12] A. Hernández-Vela, M. Reyes, V. Ponce, and S. Escalera, "GrabCut-based human segmentation in video sequences." *Sensors (Basel, Switzerland)*, vol. 12, no. 11, pp. 15 376–93, Jan. 2012.

[13] G. Gomez and E. F. Morales, "Automatic feature construction and a simple rule induction algorithm for skin detection," in *Proc. of the ICML workshop on Machine Learning in Computer Vision*, 2002, pp. 31–38.

[14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2001, pp. I–511–I–518.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition*, vol. 1, no. 3. IEEE, 2005, pp. 886–893.

[16] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning," *Foundations and Trends in Computer Graphics and Vision*, vol. 7, no. 2-3, pp. 81–227, 2011.