International Journal of Engineering & Technology, 7 (2.21) (2018) 5-8

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Interrelationship identification between humans from images using two class classifier

Amit Verma[1*], T. Meenpal[2], B. Acharya[3]

[1]*Department of Electronics and Telecommunication, NIT Raipur, G.E. Raipur, Chhattisgarh, India*
[2]*Department of Electronics and Telecommunication, NIT Raipur, G.E. Raipur, Chhattisgarh, India*
[3]*Department of Electronics and Telecommunication, NIT Raipur, G.E. Raipur, Chhattisgarh, India*
**Corresponding author E-mail:amitvermaphd@gmail.com*

Abstract

The paper proposes an automatic interrelationship identification algorithm between human beings. The image database contains two interrelationship classes i.e. two people hugging and handshaking each other. The feature detection and feature extraction has been done using bag of words algorithm. SURF features and FAST features are used as feature detectors. Finally, the extracted features have been applied to SVM for classification. We have tested the classifier against a set of test images for both feature detectors. Finally, the accuracy of the classifier has been calculated and confusion matrix has been plotted.

Keywords: *Bag of words, SVM, confusion matrix, SURF, FAST.*

1. Introduction

Interrelationship between objects in the image is an important area of research which focuses on the generation of linguistic expressions from the images usually seen by a normal human being. Identifying a relevant and useful information from the visual world around us is a challenging aspect of computer vision. Along with visual categorization of the objects, to understand the relationship between them is another important research problem. In this paper, we have proposed an algorithm to identify the interrelationship between human beings in images. We believe that identification of action will provide better description of the images.

A decent amount of work has been already done in the field of human action identification [1]using motion analysis [2], which are bounded by manual segmentation of objects and also manual detection of the actions using certain body pose analysis. The recent approaches in the field of computer vision and machine learning have provided better and autonomous solutions for these problems. We have utilized one of those possible solutions to solve our problem of interest i.e. Bag of words [3]. It corresponds to a histogram of the number of occurrences of particular image patterns in the given images. The vector representation of this particular image pattern is known as a descriptor. We have generated unique descriptors for each of the class. The process of descriptor generation [4], [5] starts with detection of interest points [6]. These interest points are used by feature detectors to extract feature. We have utilized two feature detectors i.e. SURF features [7] and FAST features [8] to extract the features around the interest points. These descriptors have been further used to train a classifier using SVM. The classifier has been tested against a set of test images to identify their classes.

In section 2 we present few important works related to visual categorization of images and identification of relationships between objects. In section 3 we explain the classification methodology in detail. In section 4 we explain the learning of the classifier using SVM. In section 5 we demonstrate the accuracy of the classifier by applying that on test and training data. Finally, in section 6 we conclude our proposed algorithm.

2. Related works

Berg et. al [9] and Bernard et. al [10] have visually categorized the faces with the names, which is comparatively easier as every face has differentiable features which can be easily detected.

Aker et al. [11] and Farhadi et. al [12] went a step ahead as they proposed a method for visual categorization of general objects. It was a difficult task, because of the variability in the visual appearance of the similar object. But it was limited to only object identification task.

Yang et al. [13] and Yao et al [14] content based image retrieval approach, in which the content of an image has been recognized first and then to construct a class for that image. But it affected the accuracy of the classification. Feng et. al [5] pro-posed another method for captioning the image using extractive generation but also assuming that corresponding object labels are available as input for the test image and the labels have to match with the objects.

Agrawal et al. [15] proposed a visual sense model using textual descriptions. He has defined the visual categorization problem as a machine learning problem and solved that by learning models i.e. SVM.

More significant work related to our approach is to detect some feature points and define a feature vectors [16], [17], [18] for the objects in the image. These feature vectors have then been utilized to learn a classifier. We have used Bag of Words to create the feature vectors and applied SVM to learn classifier.

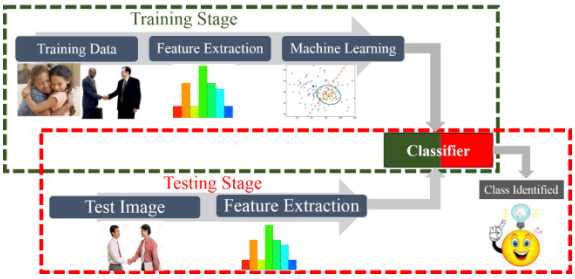

Figure 1: Block diagram of human interrelationship identification problem

3. Proposed method

The main steps of the method are:

1. Feature Point detection and Feature Extraction.
2. Generation of feature metric and code vectors.
3. Assigning these code vectors to a predefined cluster using k mean clustering to generate bag of words.
4. Learn a two-class classifier to categorize two different classes i.e. Handshaking or Hugging using SVM.

Dataset

We have taken 50 random images from internet sources belongs to handshaking class and similarly 50 images for the second class. The main advantage of using the Bag of Words models is that we do not have to resize the database images to similar pixel size as it can handle the variability of the size of the pixels. It is being assume that all the training images in the database contains only two persons with the corresponding interrelationships and with a constant background

Preprocessing

All the images in the dataset have been converted into grayscale images as we are using SURF and FAST feature points. Both of the feature point detector requires gray images as input.

Bag of words model

In this model, an image is considered as a document and the image features are called visual words. The concept of bag of feature model is very easy to understand. i.e. when we see a document; there may be certain words whose frequency is too high as compared to others. Similarly, from the training images it detects the similar patches and treat them as descriptor for those images.

Feature detection and extraction

For feature detection, we have used two different feature detectors i.e. Speed Up Robust Features (SURF) and Features from Accelerated Segment Test (FAST). The model has been tested for both the feature detectors.

SURF features [7] are robust and scale invariant feature detector with high repeatability which detects a blob like structure around the interest points from each image which is also called as SURF points. It uses Hessian Matrix approximation for blob(feature) detection. For a given point $X = (x, y)$ in an image I, the Hessian matrix $H(X, \sigma)$ in X at scale σ is defined as follows:

$$H(X,\sigma) = \begin{matrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma) & L_{yy}(X,\sigma) \end{matrix} \quad (1)$$

where $L_{xx}(X, \sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$with the image I in point X, and similarly for $L_{xy}(X, \sigma)$ and $L_{yy}(X, \sigma)$.

Once the interest point has been detected, SURF uses wavelet responses for feature extraction in horizontal and vertical direction. A neighborhood of size $20s$ is taken around the keypoint where s is the Gaussian scale at which the response of a interest point is maximum. It is divided into 4×4 sub regions. For each sub region, horizontal and vertical wavelet responses are taken and a vector is formed like this, $v = (\sum d_x, \sum d_y, |\sum d_x|, |\sum d_y|)$, where d_x is Haar wavelet response in horizontal direction and d_y is Haar wavelet response in vertical direction. This results in a feature vector for all 4×4 sub-regions of length 64.

Another feature detector which we have used here is **FAST algorithm**, which was proposed by Edward Rosten and Tom Drummond [8] which is fast enough to apply for real time applications. The FAST algorithm detects interest points also called as corner points. For a pixel C is corner if there are k connected pixel in a circle of 16 pixels around C and if all k pixel are darker than $I_c - th$ or all k pixels are brighter than $I_c + th$. where I_c is pixel intensity of C and th is some chosen threshold. Each pixel (say C) in these k pixels can have one of the following three states:

$$S_x = \begin{cases} dif I_x \leq I_c - th & (darker) \\ sI_c - th < I_x < I_c + th & (similar) \\ bI_x \geq I_c + th & (brighter) \end{cases} \quad (2)$$

where $x \in \{1,2, \ldots k\}$

Now for feature extraction process we have used the similar method as SURF which is already explained above.

These feature vectors are nothing but descriptors calculated from the neighborhood around the interest points representing local patches in the image.

$$X = [x_1 x_2 x_3 x_4 \ldots \ldots \ldots x_n] \quad (3)$$

where, X is a descriptor and collection of feature vector for each image and n shows number of local patches detected.

In each image from each classes, a set of feature points has been identified. Around these points a feature vector has been calculated. For similar points in different images, these feature vectors will produce similar values. In figure 2, few similar points identified from different images have been shown. It is very much clear that for handshaking class the fingertip is most common feature. Similarly, for the hugging class, the point of contact of the two arms or elbows of the two persons generating a unique and common feature points.

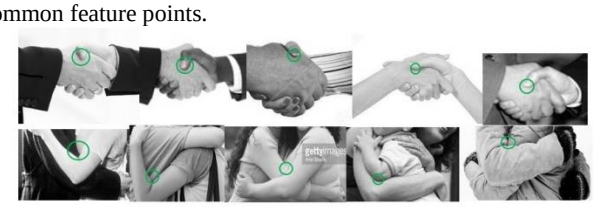

Figure 2: Important feature points detected in different images for the two classes

Generation of feature metric and codebook

Now for each feature vector a feature matric has been calculated by assigning weights to each feature vector. The stronger features are assigned more weights which helps us remove the weak features before learning. Another important reason of de-fining the feature matric is that different images may have different number of local images patches. The feature matric sort them in order and select the higher ones.

After completion of the feature metric generation, the process of codebook generation has been started. The codebook is the collection of the code vectors. The code vectors are nothing but collections of similar patches. For codebook generation, we define fix number of clusters. These clusters are nothing but the number of similar patches (code vectors) which are also called as similar words in Bag of Words analogy. Each feature matric is mapped to a code vector using K- mean clustering [19]. After mapping we have K clusters i.e. K code vectors. So, each code vector is considered as a visual word and codebook is considered as a visual dictionary as seen in figure 3. The length of each code vector represents its frequencies in the database. This codebook is

called as bag of words.

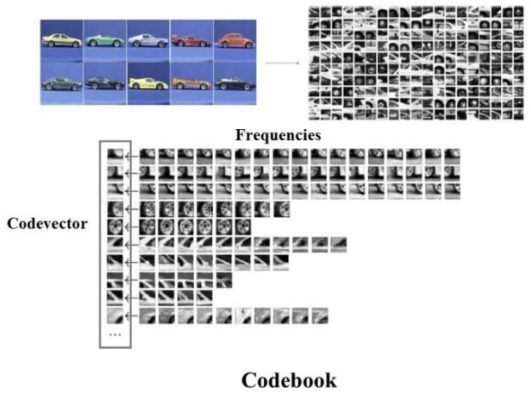

Figure 2: An example of codebook generation. Source: B. Leibe [20]

4. Learning with SVM

Once the code vectors have been assigned to the clusters we reduce our visual classification problem to multi class supervised learning problem. The classifier separates the images into two different classes i.e. Handshaking and Hugging. This classifier has been trained using SVM classifier [21]. The SVM classifier computes a hyperplane that best separates the two-class data using maximal margin approach.

$$\begin{cases} w^T x_i + b \geq 1 & for\ \forall i \quad such\ that\ y_i = +1 \\ w^T x_i + b \leq 1 & for\ \forall i \quad such\ that\ y_i = -1 \end{cases} \tag{4}$$

where feature vectors $x_i \in R^n$ and output label $y_i \in \{+1, -1\}$. w and b represents the parameters of the hyperplane. The hyperplane separates the feature matrices generated for the two classes.
As already mentioned we are using two feature detectors i.e. SURF and FAST, so we have generated two different classifiers for each feature detector. This classifier can now be used to identify a new image belongs to which class.

5. Evaluation of results

The classifiers have now been tested against the training data and test data. We have already defined 100 images for training dataset and we have also defined 100 new test images for checking the average accuracy of the classifier.
The results are shown in figures below. Figure 4 and figure 5 show the confusion matrix for training images for SURF and FAST features respectively. For SURF features out of 100 training images 97 images has been identified in correct class with average accuracy 97% whereas for FAST features out of 100 training images 95 images has been identified in correct class with average accuracy 95%. It is identified that the SURF features are giving better results than FAST features.
The results are shown in figures below. Figure 6 and figure 7 show the confusion matrix for training images for SURF and FAST features respectively. For SURF features out of 100 training images 97 images has been identified in correct class with average accuracy 97% whereas for FAST features out of 100 training images 95 images has been identified in correct class with average accuracy 95%. It is identified that the SURF features are giving better results than FAST features.
Figure 5 and figure 6 show the confusion matrix for test images for SURF and FAST features respectively. For SURF features out of 100 test images 68 images has been identified in correct class with average accuracy 68% whereas for FAST features out of 100 test images 74 images has been identified in correct class with average accuracy 74%. It is identified that the FAST features are giving better results than SURF features.

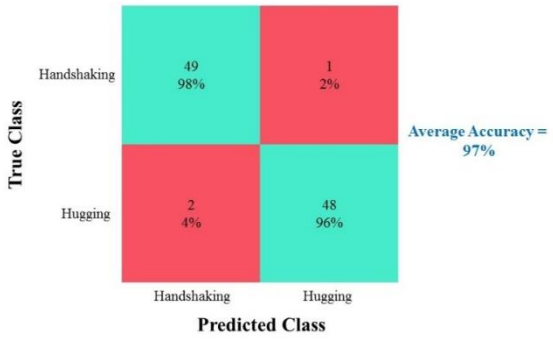

Figure 3: Confusion matrix for training images using SURF features

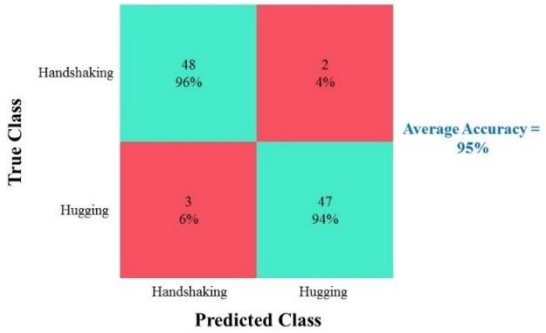

Figure 4: Confusion matrix for training images using FAST features

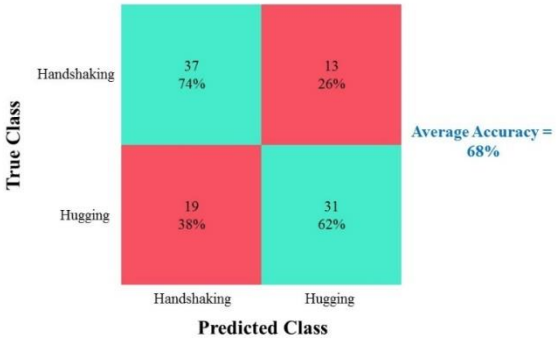

Figure 5: Confusion matrix for test images using SURF features

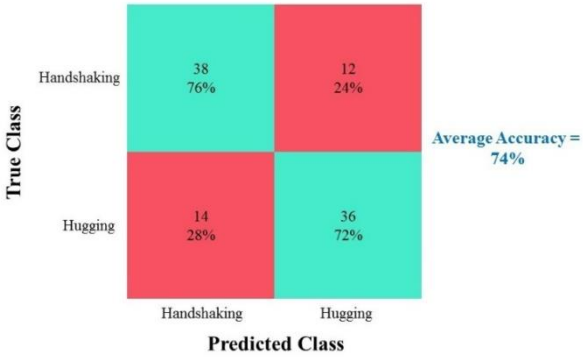

Figure 6: Confusion matrix for test images using FAST features

6. Conclusion

We have presented a simple and novel approach for interrelationship identification between two humans. Both the feature detectors i.e. SURF and FAST; are working well with the database and the two class classifier has been evaluated and producing good accuracy for both i.e. test images and training images. In near future, more number of interrelationships can be added to this problem. The classifier has now been trained using SVM which can be further extended with artificial neural networks.

References

[1] Vajda T & Marton L, "General framework for human object detection and pose estimation in video sequences", *5th IEEE International Conference on Industrial Informatics*, (2007), 467-472.

[2] Park S & Aggarwal JK, "Segmentation and tracking of interacting human body parts under occlusion and shadowing", *Proceedings Workshop on Motion and Video Computing*, (2002), pp.105-111.

[3] Csurka G, Bray C, Dance C & Fan L, "Visual categorization with bags of keypoints", *Workshop on Statistical Learning in Computer Vision, ECCV*, (2004), pp.1-22.

[4] Qin L & Gao W, "Image matching based on a local invariant descriptor", *IEEE International Conference on Image Processing*, Vol.3, (2005).

[5] Feng Y & Lapata M, "Automatic Caption Generation for News Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.35, No.4, (2013), pp.797-812.

[6] You J, Pissaloux E & Cohen HA, "A hierarchical image matching scheme based on the dynamic detection of interesting points", *International Conference on Acoustics, Speech, and Signal Processing*, Vol.4, (1995), pp.2467-2470.

[7] Bay H, Ess A, Tuytelaars T & Van Gool L, "Speeded-up robust features (surf)", *Comput. Vis. Image Underst.*, Vol.110, No.3, (2008), pp.346-359.

[8] Rosten E & Drummond T, "Machine learning for high-speed corner detection", *Proceedings of the 9th European Conference on Computer Vision-Volume Part I,ser. ECCV'06*, (2006), pp.430-443.

[9] Berg TL, Berg AC, Edwards J, Maire M, White R, Teh YW, Learned-Miller E & Forsyth DA, "Names and faces in the news", *Proceedings of the IEEE Computer Society Conferenceon Computer Vision and Pattern Recognition*, Vol.2, (2004), pp. II-848-II-854.

[10] Barnard K, Duygulu P, Forsyth D, de Freitas N, Blei DM & Jordan MI, "Matching wordsand pictures", *J. Mach. Learn. Res*., Vol. 3, (2003), pp.1107-1135.

[11] Aker A & Gaizauskas R, "Generating image descriptions using dependency relational patterns", *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ser. ACL '10*, (2010), pp.1250-1258.

[12] Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J & Forsyth D, "Every picture tells a story: Generating sentences from images", *Proceedings of the 11th European Conference on Computer Vision: Part IV, ser. ECCV'10*, (2010), pp.15-29.

[13] Yang Y, Teo CL, Daume H & Aloimonos Y, "Corpus-guided sentence generation of natural images", *Proceedings of the Conference on Empirical Methods in Natural Language Processing, ser. EMNLP '11*, (2011), pp.444-454.

[14] Yao BZ, Yang X, Lin L, Lee MW & Zhu SC, "I2t: Image parsing to text description", *Proceedings of the IEEE*, Vol.98, No.8, (2010), pp.1485-1508.

[15] Agrawal S, Verma NK, Tamrakar P & Sircar P, "Content based color image classification usingsvm", *Eighth International Conference on Information Technology: New Generations*, April (2011), pp.1090-1094.

[16] Desai C, Ramanan D & Fowlkes C, "Discriminative models for multi-class object layout", *IEEE 12th International Conference on ComputerVision*, (2009), 229-236.

[17] Gupta A & Davis LS, "Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers", *Proceedings of the 10th European Conference on Computer Vision: Part I, ser. ECCV'08*, 2008, 16-29.

[18] Torralba A, Murphy KP & Freeman WT, "Using the forest to see the trees: exploiting context for visual object detection and localization", *Communications of the ACM*, Vol.53, No.3, (2010), pp.107-114.

[19] Boser BE, Guyon IM & Vapnik VN, "Atraining algorithm for optimal margin classifiers", *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, ser. COLT '92*, (1992), pp.144-152.

[20] Leibe B, Leonardis A & Schiele B, "Robustobject detection with interleaved categorization and segmentation", *Int. J. Comput. Vision*, Vol.77, No.1-3, (2008), pp.259-289.

[21] Weston J & Watkins C, "Multi-class support vector machines", *Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London*, (1998).