On the Creation of Visual Models for Keywords through Crowdsourcing

ZENONAS THEODOSIOU, NICOLAS TSAPATSOULIS
Cyprus University of Technology
Department of Communication and Internet Studies
31 Archbishop Kyprianos Str, CY-3036, Limassol
CYPRUS
{zenonas.theodosiou, nicolas.tsapatsoulis}@cut.ac.cy

Abstract: Crowdsourcing annotation is a recent development since a complete and elaborate annotation of the content of an image is an extremely labour-intensive and time consuming task. In this paper we examine the possibility to build accurate visual models for keywords created through crowdsourcing. Specifically, 8 different keywords related to athletics domain have been modelled using MPEG-7 and Histogram of Oriented Gradients (HOG) low level features and the Sequential Minimal Optimization (SMO) classifier. The experimental results have been examined using accuracy metrics and are very promising showing the ability of the visual models to classify the images into the 8 classes with the highest average accuracy rate of 73.13% in the purpose of the HOG features.

Key–Words: Crowdsourced Annotation, Visual Models, Low level Features

1 Introduction

The rapid growth of digital image libraries creates the need of effective image tagging. Manual image annotation is an extremely difficult and elaborate task and cannot always be considered as correct due to visual information that always lets the possibility for more individual interpretation and ambiguity [1]. Automatic image annotation is currently an important research topic in the field of computer vision [2] and attempts to learn the afore-mentioned correlation and build a dictionary between low-level features and high-level semantics [3]. A manually annotated set of multimedia data is used to train a system for the identification of joint or conditional probability of an annotation occurring together with a certain distribution of multimedia content feature vectors [4]. Different models and machine learning techniques are developed to learn the correlation between image features and textual words from the examples of annotated images and then apply the learned correlation to predict words for unseen images [5].

Multiple judgements per image from several annotators can partially solve the problem of semantic annotation multimedia data and improve the annotation quality. The act of outsourcing work to a large crowd of workers is rapidly changing the way datasets are created [6]. The fact that differences between implicit and explicit relevance judgments are not so far [7] opened a new way, where implicit relevance judgments were considered as training data for various machine learning-based improvements to information retrieval [8], [9]. In the current study we investigate the possibility of creating visual models for crowdsourced annotations using two different types of low level features and the SMO classifier. For the performance of the proposed method, fifteen users annotated a set of a 500 images taken from the athletics domain, using a predefined set of keywords. Images sharing a common keyword are grouped together and used for creating the visual model which corresponds to this keyword. We have used publicly available tools for the computation of the low level features [10], [11] and the model creation (the Weka tool [12]) and classified the images into 8 keyword classes.

The remaining paper is organized as follows: Section 2 presents the method we have followed to create the dataset and model the keywords while Section 3 gives a detailed description of the used low level features. Experimental results and discussion are reported in Section 3. Finally, conclusions are drawn and further work hints are given in Section 4.

2 Method Overview

2.1 Dataset Creation & Keyword Modelling

A randomly selected set of 500 images taken from a large dataset created during the FP6 BOEMIE project

ISBN: 978-1-61804-074-9

```
- <root>
  - <IndoorEvent>
    - <Jumping>
      + <HighJump>
      - <PoleVault>
          <Athlete />
          <Pillars />
          <HorizontalBar />
          <Pole />
        </PoleVault>
      + <LongJump>
      + <TripleJump>
      </Jumping>
    - <Throwing>
      + <DiscusThrow>
      + <HammerThrow>
      + <JavelinThrow>
      + <ShotThrow>
      </Throwing>
    + <Running>
    </IndoorEvent>
  - <OutdoorEvent>
    + <Marathon>
    + <Walking>
    </OutdoorEvent>
  </root>
```

Figure 1: The XML dictionary used for annotation.

was manually annotated by fifteen users using the MuLVAT annotation tool [13] with the aid of a structured xml dictionary (Figure 1). An overview of the dataset creation and keyword modelling procedure is provided in Figure 2. For our experiments we have selected a set of 8 representative keywords from a total number of 33 dictionary keywords. The set of the selected keywords is represented by $\boldsymbol{K} = \{K_1, ..., K_N\}$. The K_i indicates the *i-th* keyword while the total number of keywords is denoted by *N*. The 8 vocabulary keywords used for our experimental setup are: (1) "Discus", (2) "Hammer", (3) "High Jump", (4) "Hurdles", (5) "Javelin", (6) "Long Jump", (7) "Running", and (8) "Triple Jump". For each keyword, we selected 50 images which were annotated from more than 5 annotators with this keyword. Examples of the selected images are given in Figure 3. The images for each keyword are grouped together and create the set of groups $\boldsymbol{G} = \{G_1, ..., G_N\}$, where G_i denotes the *i-th* group of a total number of *N* groups of images. The MPEG-7 and HOG features were extracted and used to create the visual models $\boldsymbol{V} = \{V_1, ..., V_N\}$, where V_i indicates the visual model for the keyword K_i.

To facilitate effective and efficient learning, each keyword is treated as a separate binary classification problem. We have followed the one-against-rest approach [14] and we have built a total number of *N* models, one for each keyword. The feature vectors of each keyword class were split into two groups, called the training (80%) and testing (20%) set. Each model

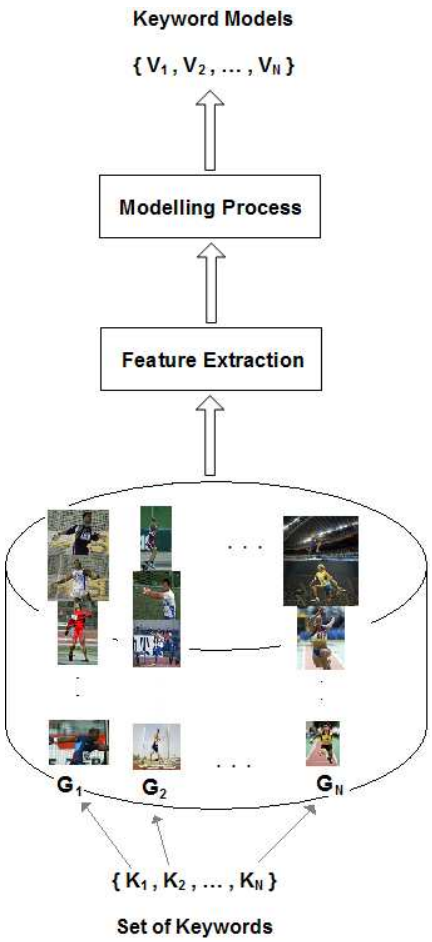
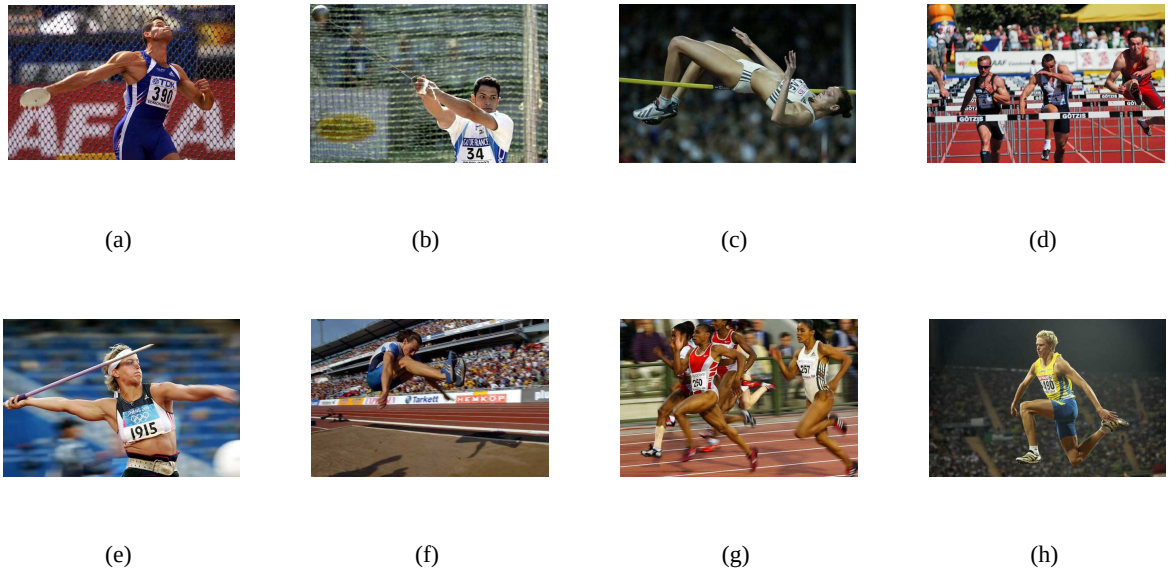


Figure 2: An overview of the dataset creation and keyword modelling.

is trained and tested between one class and the *N*-1 other classes. The training and testing set for each model contain the feature vectors of the corresponding keyword class and the same number of randomly selected feature vectors of the the rest *N*-1 classes. Keywords models were created using Weka tool [12]. Among a variety of possible classifiers we decided to use one of the state of the art implementations of the Support Vector Machines (SVMs), the Sequential Minimal Optimization (SMO) [15], [16]. It has been reported in several publications as the best performing machine learning algorithm for a variety of classification tasks. The performance of SMO classifiers can vary significantly with variation in parameters of the models. During training we experimented with different parameters and kernels and for each kernel we built models for several combinations of the parameters, with the Puk kernel performing better than the



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

(e) (f) (g) (h)

Figure 3: Images from the athletics domain corresponding to the following classes: (a) Discus, (b) Hammer, (c) High Jump, (d) Hurdles, (e) Javelin, (f) Long Jump, (g) Running, (h) Triple Jump.

others.

3 Low Level Feature Extraction

Among the possible low level features that can be extracted from an image, we have chosen to use and compare the MPEG-7 and HOG.

3.1 MPEG-7 features

MPEG-7 visual descriptors include the color, texture and shape descriptor. A total of 22 different features are included, nine for color, eight for texture and five for shape. The dominant color features include color value, percentage and variance and require especially designed metrics for similarity matching. Furthermore, their length is not known a priori since they are image dependent (for example an image may be composed from a single color whereas others vary in color distribution). The previously mentioned difficulties cannot be easily handled in machine learning schemes, therefore we decided to exclude these features for the current experimentation. The texture browsing features (regularity, direction, scale) have not been included in the description vectors since in the current implementation of the MPEG-7 experimentation model [10] the corresponding descriptor cannot be reliably computed (it is a known bug of the implementation software). The scalable color and shape descriptor features have been also excluded because vary depending on the form of an input object and can not be used for the holistic image description. Among all MPEG-7 descriptors only the Color Layout (CL), Color Structure (CS), Edge Histogram (EH) and Homogenous Texture (HT) descriptors are used in our experiments.

3.2 Histogram of Gradients Features

The HOG features exploit the idea that local object appearance can be described by the distribution of intensity gradients or edge directions. The image is divided into small connected regions, called cells. For each cell, a histogram of gradient directions or edge orientations within this cell is compiled. For the implementation of HOG, each pixel within the cell casts a weighted vote for an orientation-based histogram channel. For the current study we have used the implementation proposed in [11] with the aid of 25 rectangular cells and 9 bins histogram per cell. The 25 histograms with 9 bins were then concatenated to make a 225-dimensional feature vector.

4 Experimental Results and Discussion

We used the dataset and keyword modelling process described in Section 2 to examine the performance and effectiveness of the created visual models. The

ISBN: 978-1-61804-074-9

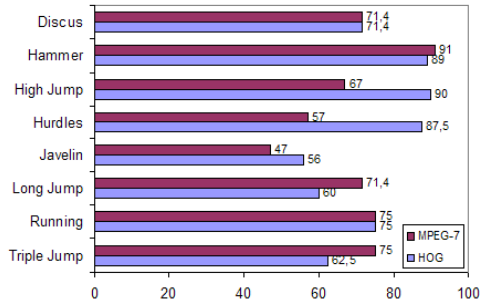


Figure 4: TPR (%) per class using MPEG-7 and HOG features.

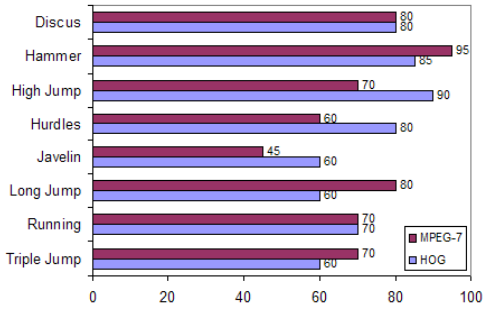


Figure 5: ACC (%) per class using MPEG-7 and HOG features.

basic aim of our experimental study was to investigate if crowdsourcing annotations can be used to create visual models. In addition, we compared the classification efficiency using the MPEG-7 and HOG features. We considered two metrics to estimate the effectiveness of the created visual models: the Total Positive Rate (TPR), which indicates the accuracy of correctly classified instances in the corresponding classes and, the Accuracy Rate (ACC), which is defined as the sum of true positives and true negatives divided by the total number of instances. For the models created using MPEG-7 features the TPR metric gave an average value of 69.35% and the average ACC metric had value of 71.25%. The corresponding values for the models created using HOG were 73.93% and 73.13%, respectively. As a consequence, the average Error Rate (ERR), which is defined as the incorrectly classified instances, is less than 30% for all experiments. Figures 4 and 5 summarize the TPR and ACC metrics for all classes using MPEG-7 and HOG features. Although nearly all models are able to classify the images into the corresponding classes, the worst efficiency is perceived when testing the "Javelin" model. This may happens because the content of images belong to "Javelin" has many similarities with the content of images belong to other keywords. Considering the classification efficiency of the low level features, it is evident from Table 1 that HOG are more reliable than MPEG-7 having the highest values for average TPR and ACC metrics.

Table 1: Average values for the used classification metrics.

Features	***TPR** (%)*	***ACC**(%)*	***ERR**(%)*
MPEG-7	69.35	71.25	28.75
HOG	73.93	73.13	26.87

5 Conclusions and Future Work

In the current study we tried to model the view of several annotators on tagging images related to the athletics domain. Specifically, 8 different keywords were modelled using low level features and the SMO classifier with the aid of the Weka tool. The experimental results show that nearly all created models can accurately classify the images into the 8 classes. There is a significant variation on the efficiency of the various features with the HOG having the highest performance. Our future work includes the investigation on larger and different datasets, experimentation of additional training algorithms and other classifications schemes. In addition, the performance accuracy of more low level features will be examined.

Acknowledgements: This work falls under the Cyprus Research Promotion Foundation's Framework Programme for Research, Technological Development and Innovation 2009-2010 (DESMI 2009-2010), co-funded by the Republic of Cyprus and the European Regional Development Fund, and specifically under Grant ΠΕΝΕΚ/0609/95.

References:

[1] T. Volker, A. Thom, S. M. M. Tahaghoghi, *Modelling human judgment of digital imagery for multimedia retrieval*, IEEE Transactions on Multimedia, vol. 9(5), pp. 967-974, 2007.

[2] A. Hanbury, *A survey of methods for image annotation*, Journal of Visual Languages and Computing, vol.19, pp. 617-627, 2008.

[3] P. Duygulu, K. Barnard, J. F. K. de Freitas, D. A. Forsyth, *Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary*, In. A. Heyden, G. Sparr, M. Nielsen, P. Johansen (eds.) ECCV 2002.

LNCS, vol.2353, pp.97-112, Springer, Heidelberg, 2002.

[4] K. Athanasakos, V. Stathopoulos, J. Jose, *A Framework for Evaluating Automatic Image Annotation Algorithms*, Advances in Information Retrieval, Lecture Notes in Computer Science, vol. 5993, pp. 217-228, 2010.

[5] R. Zhang, Z. Zhang, M. Li, W. Y. Ma, H. J. Zhang, *A probabilistic semantic model for image annotation and multi-modal image retrieval*, Multimedia Systems, vol.12(1) pp.27-33, 2006.

[6] P. Welinder, P. Perona, *Online crowdsourcing: rating annotators and obtaining cost effective labels*, In Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 25-32, 2010.

[7] T. Joachims, L. Granka, B. Pang, H. Hembrooke, G. Gay, *Accurately interpreting clickthrough data as implicit feedback*, In Proc. of the 28th Annual International ACM SIGIR Conference, pp. 154-161, 2005.

[8] C. Macdonald, I. Ounis, *Usefulness of quality click-through data for training*, In Proc. of the 2009 Workshop on Web Search Click Data, pp. 75-79, 2009.

[9] T. Tsikrika, C. Diou, A. P. de Vries, A. Delopoulos, *Image annotation using clickthrough data*, In Proc. of the 8th International Conference on Image and Video Retrieval, 2009.

[10] MPEG-7 Visual Experimentation Model (XM), Version 10.0, ISO/IEC/JTC1/SC29/WG11, Doc. N4063, 2001.

[11] O. Ludwig, D. Delgado, V. Goncalves, U. Nunes, *Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection*,In Proc. of 12th International IEEE Conference On Intelligent Transportation Systems, V. 1., pp. 432-437, 2009.

[12] I. H. Witten, E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[13] Z. Theodosiou, A. Kounoudes, N. Tsapatsoulis, M. Milis, *MuLVAT: A Video Annotation Tool Based on XML-Dictionaries and Shot Clustering*, In Proc. of the 19th International Conference on Artificial Neural Networks: Part II, pp. 913-922, 2009.

[14] D. M. J. Tax, R .P .W. Duin, *Using two-class classifiers for multiclass classification*, In Proc. of 16-th International conference of Pattern Recognition, pp.124-127, 2002.

[15] J. Platt, *Fast Training of Support Vector Machines using Sequential Minimal Optimization*, Advances in Kernel Methods-Support Vector Learning, B. Schoelkopf, C. Burges, and A. Smola, eds., MIT Press, 1998.

[16] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, K. R. K. Murthy, *Improvements to Platt's SMO Algorithm for SVM Classifier Design*, Neural Computation, vol. 13(3), pp.637-649, 2001.