Frame the Crowd: Global Visual Features Labeling boosted with Crowdsourcing Information

Michael Riegler
Klagenfurt University
Klagenfurt, Austria
miriegle@edu.uni-klu.ac.at

Mathias Lux
Klagenfurt University
Klagenfurt, Austria
mlux@itec.uni-klu.ac.at

Christoph Kofler
Delft University of Technology
Delft, The Netherlands
c.kofler@tudelft.nl

ABSTRACT

In this paper we present our approach to the Crowd Sourcing Task of the MediaEval 2013 Benchmark [2] using transfer learning and visual features. For the visual features we adopt an existing approach for *search based classification* using content based image retrieval on global features with *feature selection* and *feature combination* to boost the performance. Our approach gives a baseline evaluation indicating the usefulness of global visual features, hashing and search-based classification.

1. INTRODUCTION

The benchmarking task at hand [2] has been investigated by two different means as well as a combination of them. First, we only use crowdsourcing data for the labeling. We compute a reliability measure for workers and use this value along with the workers' self-reported familiarity as features for a classifier. Our second approach is based on the assumption, that images taken with similar intentions, i.e. displaying a fashion style, are *framed* in a similar way.

We define the *framing of an image* as the sum of the visible reflexes of the specific decisions that the photographer makes when the image is captured. The photographer has many different choices when taking a photo of a certain object, event, person or scene. During the capture process the photographer does not click the shutter randomly, but rather makes use, either consciously or unconsciously, of a set of conventions that can be thought of as a recipe for a certain kind of image. The recipe leads to a distinguishable framing that is used by the viewer in interpreting the image. For example, a picture of a person framed in one way is most easily interpreted as a fashion image and framed in another way most easily interpreted as a holiday memory. Choices photographers make to achieve certain types of framing include color distribution, lighting, positions of objects and people etc. They also include the choice of the exact moment during ongoing action at which the image is shot. In this way, the photographer also influences exactly what is depicted in the image, e.g., facial expressions of the people appearing in the image. Especially for fashion use cases the framing theory is applicable. Due to the nature of framing we employ global visual features using and modifying the LIRE framework [4] and boost classification results with feature combination and feature selection.

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

2. APPROACH

Using LIRE we extracted the global features CEDD, FCTH, JCD, PHOG, EH, CL, Gabor, Tamura, LL, OH, JPEGCoeff and SC (which are described and referenced in [4]). These features are able to detect and distinguish characteristics of a framing like the color distribution with Color Layout.

The task includes one required condition, which only allows the use of the workers' annotations. However, it is noted that those annotations are error prone. Therefore, we integrated a reliability measure for workers based on the work of Ipeirotis et al. [1].

We compare a rating $r_{ij} \in R$ from worker $w_i \in W$ for image $x_j \in I$ with R being the set of ratings, W being the set of workers and I being the set of images, to the majority votes $V(x_j)$ of all workers for an image. This gives a measure of reliability $Q(w_i)$ of a specific worker w_i. The computed weight $Q(w_i)$ is then multiplied with the vote r_{ij} of the worker w_i for the image x_j.

$$V(x_j) = \underset{v \in \{0,1\}}{\arg\max} \mid \{r_{ij} : w_i \in W \wedge r_{ij} = v\} \mid$$

$$Q(w_i) = \frac{\mid \{r_{ij} : x_j \in I \wedge r_{ij} = V(x_j)\} \mid}{\mid \{r_{ij} : x_j \in I\} \mid}$$

Additionally, the familiarity of the worker with the fashion topic is also added as a feature. So the feature vector for an image i_k with ratings of three workers w_1, w_2, w_3 is $(r_{1k} \cdot Q(w_1), r_{2k} \cdot Q(w_2), r_{3k} \cdot Q(w_3), f_{w1}, f_{w2}, f_{w3})$.

3. EXPERIMENTS

We submitted five different evaluation runs. For the crowdsourcing task, a two part data set was available. The first part (MMSys dataset) is described in [3] and will be referenced, for convenience, in this paper as D_M. The second part of the data is called the Fashion 10000 data set (D_F). To transfer the experts' knowledge from D_M to D_F, we use a process called *transfer learning* for all our runs. This is done by using a model, built from an expert knowledge containing data set (in this case D_M) to generate a new accurate model for the dataset without expert knowledge. In this case, by labeling the images from D_F with the D_M model.

The first evaluation run – the required one, run #1 – made use of the feature vector of worker annotations, mentioned above, and the Weka *Random Forest Classifier*, which yielded good results in cross validation on D_M. Using a model built from the D_M data set we labeled the images from D_F and retrained our model using the newly labeled images.

For the visual content classification (run #2) we used D_M to build a model for classification. The classifier is search based, which means that the image being classified is considered as query and the label is derived from the result list. A similar approach was used in [5]. Each result in the list *votes* for a label, weighted by its inverse rank. The selection of global features for classification is based on the information gain of each global feature with respect to the class labels in the training set D_M. For the combination, only features that have an above-average information gain are used. The combination of the global features is done with *late fusion*. This means, each global feature has its own classifier and returns a ranked list for the given query image. Label weights (inverse rank) are than added up, resulting in a combination by rank.

Classification performance in terms of time and scaling is promising. In the worst case with 12 features combined, classification per image takes about 240 ms. In the best case – if only one feature is used – classification time is down to 16 ms per image.

Run #3 uses the same techniques as described for run #2, but uses the worker annotations of D_F for training the model. Run #4 uses the images labeled in run #1 for training, and run #5 combines run #1 and run #4 in a way, that classification based on visual features is used when the random forest classifier returns an uncertain result.

4. DISCUSSION AND CONCLUSIONS

To estimate the performance of each run we used the test data set and D_M experts votes for ground truth. We split the dataset 80% for training and 20% for test. The results of these tests can be seen in Table 1 for both labels (L_1, L_2). L_1 stands for whether an image is fashion related or not and L_2 stands for whether the content of the image matches with the category for the fashion item depicted in the image. For the evaluation we used weighted F1 score (WF1), because the positive and negative classes are not comparable on size. The results of the Benchmark can be seen in Table 2. The tests results show that the crowd sourcing classifier has the best performance. Also the official results support this fact. The outcome for the workers information based runs in the final results compared to our test results indicates that transfer learning worked well.

Visual features based on classification performs much better in our tests than in the final results (cp. runs #2-#4). It's common that metadata, even when generated by crowdsourcing, leads to better results, but still the performance drop between preliminary and official results is obvious. However, WF1 scores are more suitable for a steady judgment as shown in Table 1 (e.g. run #3, F1 vs. WF1 scores in the preliminary runs).

Nevertheless, taking all constraints into account, the visual features perform quite well. Their benefit is that, unlike crowdsourcing, which costs money, the effort to extract them and get a small amount of training data is minimal. Moreover, metadata quality depends on the actual workers and the quality control mechanism of the crowdsourcing platform. This is also indicated by the lower WF1 measure of run #3 compared to run #2, as in run #2 expert votings were used to train the model, while run #3 also takes crowdsourcing workers into account for training.

Another interesting effect is that a combination of crowdsourcing metadata and visual content can improve the performance, if it is used to build the model of the visual classifier. In the other direction, it seems to lower performance. Visual information based models have already worked well with small number of training data and a small amount of crowdsourcing can help to boost visual information retrieval systems performance.

Table 1: Preliminary Test Results

Run	F1 L_1	F1 L_2	WF1 L_1	WF1 L_2
1	0.882	0.882	0.872	0.915
2	0.7669	0.2599	0.7368	0.6047
3	0.7483	0.0493	0.623	0.5215
4	0.7608	0.1204	0.6894	0.5489
5	0.885	0.892	0.883	0.932

Table 2: Official MediaEval Results

Run	F1 L_1	F1 L_2
1	0.7124	0.7071
2	0.5201	0.2908
3	0.4986	0.4269
4	0.5403	0.3938
5	0.7123	0.6999

We further assume that our theory of framing is supported by the results. Especially our test results, because Label 1 is very good detectable by our global features classifier. On the other side, the Label 1 detection was not good. This is logical, because for the task of object detection local features are better suitable.

For future work it will be interesting to take a closer look on the relationship between crowdsourcing and how it could be used to improve the performance of visual features and vice versa. Another interesting direction would be to use crowdsourcing to create a specific dataset for framing. This would help to draw a clearer definition and show the usefulness of the framing theory.

5. REFERENCES

[1] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM, 2010.

[2] B. Loni, A. Bozzon, M. Larson, and L. Gottlieb. Crowdsourcing for Social Multimedia at MediaEval 2013: Challenges, data set, and evaluation. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[3] B. Loni, M. Menendez, M. Georgescu, L. Galli, C. Massari, I. S. Altingovde, D. Martinenghi, M. Melenhorst, R. Vliegendhart, and M. Larson. Fashion-focused creative commons social dataset. In *Proceedings of the 4th ACM Multimedia Systems Conference*, MMSys '13, pages 72–77, New York, NY, USA, 2013. ACM.

[4] M. Lux. LIRE: Open source image retrieval in java. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, page to appear, New York, NY, USA, 2013. ACM.

[5] L. Yang and A. Hanjalic. Supervised reranking for web image search. In *Proceedings of the international conference on Multimedia*, pages 183–192. ACM, 2010.