Rocchio-Based Relevance Feedback in Video Event Retrieval

G.L.J. Pingen[1,2(✉)], M.H.T. de Boer[2,3], and R.B.N. Aly[1]

[1] University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
r.aly@utwente.nl
[2] TNO, P.O. Box 96864, 2509 JG The Hague, The Netherlands
{geert.pingen,maaike.deboer}@tno.nl
[3] University of Nijmegen, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands

Abstract. This paper investigates methods for user and pseudo relevance feedback in video event retrieval. Existing feedback methods achieve strong performance but adjust the ranking based on few individual examples. We propose a relevance feedback algorithm (ARF) derived from the Rocchio method, which is a theoretically founded algorithm in textual retrieval. ARF updates the weights in the ranking function based on the centroids of the relevant and non-relevant examples. Additionally, relevance feedback algorithms are often only evaluated by a single feedback mode (user feedback or pseudo feedback). Hence, a minor contribution of this paper is to evaluate feedback algorithms using a larger number of feedback modes. Our experiments use TRECVID Multimedia Event Detection collections. We show that ARF performs significantly better in terms of Mean Average Precision, robustness, subjective user evaluation, and run time compared to the state-of-the-art.

Keywords: Information retrieval · Relevance feedback · Video search · Rocchio · ARF

1 Introduction

Finding occurrences of events in videos has many applications ranging from entertainment to surveillance. A popular way to retrieve video events from a given collection is to combine detection scores of related concepts in a ranking function. However, selecting related concepts and defining query specific ranking functions without any examples is challenging. Text retrieval, which is the corresponding problem in the text domain, strongly benefits from relevance feedback, which defines ranking functions based on a set of examples with assumed relevance status [15]. While the basic relevance feedback principle recently also gained popularity in video retrieval, the methods of adapting ranking functions are generally developed from scratch. However, from a scientific standpoint it would be more desirable to transfer the knowledge gained for text retrieval into approaches in video retrieval, especially if those methods show stronger performance. Therefore, this paper derives a novel concept-based

L. Amsaleg et al. (Eds.): MMM 2017, Part II, LNCS 10133, pp. 318–330, 2017.
DOI: 10.1007/978-3-319-51814-5_27

relevance feedback algorithm that is derived from a proven relevance feedback algorithm in text retrieval.

The state-of-the-art relevance feedback algorithms in video retrieval update the ranking function based on few examples of the positive and the negative class [5,6]. However, especially in pseudo relevance feedback, which assumes a set of examples to be relevant without actual feedback, we find this hurts performance for some queries because these assumptions are not met. In text retrieval, on the other hand, relevance feedback algorithms update the ranking function based on the centroids in the relevant and non-relevant examples. We propose that this approach is likely to be more robust, because considering centroids evens out outliers in the examples. Furthermore, adapting ranking functions based on the differences in the centroids is also more likely to generalize to the remaining videos in the collection, potentially improving effectiveness overall.

This paper evaluates the proposed relevance feedback algorithm based on several real and pseudo user feedback modes and investigates whether a different trend between modes of relevance feedback exist.

In the next section, we present related work on relevance feedback. The third section explains our Adaptive Video Event Search system (AVES), which uses the proposed relevance feedback algorithm. The fourth section contains the experimental set-up using the TRECVID Multimedia Event Detection benchmark and the fifth section consists of the results of both simulation and user experiments. We end this paper with a discussion and the conclusions and future work.

2 Related Work

Relevance feedback in video event retrieval is an increasingly active field of research. In video retrieval, Dalton et al. [3] showed that pseudo-relevance feedback can increase Mean Average Precision up to 25%, and with human judgment this number can grow up to 55%. The state-of-the-art methods, such as feature-, navigation-pattern, and cluster-based methods, in image retrieval are explained by Zhou et al. [25] and Patil et al. [14]. Oftentimes the system will actively select the documents that achieve the maximal information gain [18]. Other methods use decision trees, SVM's, or multi-instant approaches are explained in Crucianu et al. [2]. Xu et al. [20] present an interactive content-based video system that incrementally refines the user's query through relevance feedback and model visualization. Their system allows a user to select a subset of relevant retrieved videos and use those as input for their SVM-based video-to-video search model, which has been shown to outperform standard text-to-video models. Yang et al. [21] also introduce an SVM based supervised learning system that uses a learning-to-rerank framework in combination with an adapted reranking SVM algorithm. Tao et al. [17] improves on the SVM-based methods using orthogonal complement component analysis (OCCA). According to Wang et al. [19], SVM-based RF approaches have two drawbacks: (1) multiple feedback interactions are necessary because of the poor adaptability, flexibility and robustness of the original visual features; (2) positive and negative samples are treated equally, whereas

the positive and negative examples provided by the relevance feedback often have distinctive properties. Within the pseudo relevance feedback, this second point is taken by Jiang et al. [7–9], who use an unsupervised learning approach in which the 'easy' samples are used to learn first and then the 'harder' examples are increasingly added. The authors define the easy samples as the videos that are ranked highest and have smaller loss. The more easy samples are those that are presumably more relevant, and would be ranked higher than others. The system then iterates towards more complex (lower-ranked) videos. An SVM/Logistic-regression model is trained using pseudo labels initiated with logarithmic inverse ordering. This approach reduces the cost of adjusting the model too much when learning from data that is very dissimilar from the learned model. Experiments show that it outperforms plain retrieval without reranking, and that has decent improvements over other reranking systems.

Another field of study in which relevance feedback is often used is in text retrieval. One of the most well-known and applied relevance feedback algorithms that has its origins in text-retrieval is the Rocchio algorithm [15]. This algorithm works on a vector space model in which the query drifts away from the negatively annotated documents and converges to the positively annotated documents. The Rocchio algorithm is effective in relevance feedback, fast to use and easy to implement. The disadvantages of the method are the parameters that have to be tuned and it cannot handle multimodal classes properly.

Another vector space model uses a k-NN method and is used by Gia et al. [6] and Deselaers et al. [5]. k-NN based methods are shown to be effective, are non-parametric, but run time is slower and it can be very inaccurate when the training set is small.

3 Adaptive Video Event Search

In this section, we explain our Adaptive Video Event Search system, named AVES. All relevant feedback algorithms depend on an *initial ranking*. A user can retrieve this initial ranking by entering a textual query into our search engine. The initial score s_v by which a video v is ranked is defined as:

$$s_v = \sum_{d \in D} w_d \cdot (s_{v,d} - b_d), \tag{1}$$

where d is the concept detector, D is set of selected concept detectors, w_d is the weight of concept detector d, $s_{v,d}$ is the concept detector score for video v and b_d is the average score on the background dataset of concept detector d.

Concept detectors are models that are trained to detect concepts in images or videos based on machine learning techniques, such as neural networks or SVMs. For the definition of D and the initial setting of w_d, we adapt a method proposed in Zhang et al. [23], first comparing the query to each of the pre-trained concepts available to our system. The skip-gram negative sampling Word2vec model from Milokov et al. [12] is used in this comparison. The pre-trained GoogleNews model, which is trained on one billion words, embeds the words into 300 dimensions.

The cosine similarity is used to calculate distances between words. The distance between the label of the concept detector and the user query is used as a weight. The thirty concepts with the highest similarities and a value higher than a threshold of 0.35 keep their weights and the other concepts have a weight of zero. In experiments outside of the scope of this paper we found that including a concept detector background score as prior is beneficial to ranking accuracy.

3.1 Adaptive Relevance Feedback (ARF)

In this section, we explain our relevance feedback algorithm, named *ARF*. This method is inspired by the Rocchio algorithm [15]. Different from other algorithms, we use relevance feedback to update the weights for our concept detectors. By updating the weights using relevance feedback our algorithm is more robust to few or wrong annotations. In k-NN methods, wrong annotations can have a high impact on ranking performance. By taking into account the initial concept detector cosine distance to the query, the proposed algorithm is more robust to this type of relevance feedback.

The weights are updated using the following formula:

$$\begin{aligned} w'_d &= w_d + (\alpha \cdot m_R) - (\beta \cdot m_{NR}) \\ m_R &= \frac{\sum_{v \in R} s_{v,d} - b_d}{|R|} \\ m_{NR} &= \frac{\sum_{v \in NR} s_{v,d} - b_d}{|NR|}, \end{aligned} \tag{2}$$

where v is the considered video, d is the concept detector, R is the set of relevant videos, NR is the set of non-relevant videos, $s_{v,d}$ is the score for concept detector d for video v, w_d is Word2vec similarity between the concept detector d and the query, b_d is the average score on the background dataset of concept detector d, and α and β are Rocchio weighting parameters for the relevant and non-relevant examples respectively.

The adjusted concept detector weight, w'_d, is then plugged back into the ranking function (see 1), where we substitute the original Word2vec score for the adjusted weight. This results in new scores, s'_v, for each video v, which is used to create an updated ranked list of videos.

3.2 Experimental Set-up

We use the MEDTRAIN set (5594 videos, 75.45 keyframes on average) and MEDTEST set (27276 videos, on average 57.11 keyframes) from the TRECVID Multimedia Event Detection benchmark [13] as our evaluation datasets. The datasets from this international benchmark is used because of the challenging events, many videos and wide acceptance. The MEDTRAIN contains relevance judgments for forty events, whereas MEDTEST only contains judgments for twenty events but this set is often used in other papers to report performance

on. Different from the MEDTRAIN in which we use the top 30 concept detectors, in the MEDTEST we only use the top 5 concept detectors, because of the bigger imbalance of positive and negative videos in the MEDTEST (20:27276) compared to the MEDTRAIN (100:5494). This imbalanced caused more concepts to a very low initial performance, whereas only 5 concepts had less noise and, therefore, slightly higher performance.

In the MEDTRAIN set only thirty-two events are used, because of the correlation of certain concept detectors to a certain event, resulting in a (near-)perfect retrieval result. Since such retrieval results are not interesting for relevance feedback application purposes, we do not consider them for the purpose of this experiment. The omitted events are the following: *Wedding ceremony*; *Birthday party*; *Making a sandwich*; *Hiking*; *Dog show*; *Town hall meeting*; *Beekeeping*; *Tuning a musical instrument*. In additional subjective user evaluation experiments, we verified that the trends obtained with the thirty-two events are also present with the forthy events, only the MAP is slightly higher.

The BACKGROUND set (5000 videos) from the benchmark is used to obtain the background scores b_d in Eq. 1. A total of 2048 concept detectors (D) were used from the ImageNet (1000) [4], Places (205) [24], SIN (346) [13] and TRECVID MED dataset (497) [13]. The concept detectors from the TRECVID MED are manually annotated on the Research set, comparable to Natarajan et al. [22] and Zhang et al. [23]. The output of the eight layer of the DCNN network trained on the ILSVRC-2012 [4] is used as concept detector score per keyframe/image. This DCNN architecture is fine-tuned on the data in the dataset for SIN, Places and TRECVID MED. For each video in our dataset we have extracted 1 keyframe per 2 s uniformly from a video. We use max pooling to obtain a concept detector score per video. We purposely did not use higher level concept detectors, such as those available in the FCVID [10] or Sports [11] dataset, to obtain more interesting experiments using relevance feedback. We, therefore, do not aim at highest possible initial ranking, but at a gain with the use of relevance feedback. We believe this is applicable to real world cases, because relevant high level concepts are not always present.

The ARF parameters α and β were taken to be 1.0 and 0.5, in line with text-information retrieval literature [15]. Variations of these parameters were also investigated, as shown in Sect. 3.3.

We compare our relevance feedback algorithm to (1) a baseline without relevance feedback *Initial*; and (2) a k-NN based relevance feedback algorithm named *RS*. The RS algorithm is well-performing in image retrieval [5,6] and the relevance score $relevance(v)$ of a video v calculated as

$$relevance(v) = \left(1 + \frac{dR(v)}{dNR(v)}\right)^{-1}, \tag{3}$$

where dR is the dissimilarity, measured as Euclidean distance, from nearest video in relevant video set R, dNR is the dissimilarity from nearest video in non-relevant video set NR.

The SVM-based methods are not included in this paper, because preliminary experiments showed that on average performance is poor due to limited amount of positive samples.

Modes. We compared the algorithms with respect to the following feedback modes: (1) the mode *Optimal* uses the ground truth to select all relevant videos, (2) the *Pseudo* relevance feedback mode selects the first 10 videos as positive (and the rest as negative), (3) the *Random* mode selects 10 positive videos at random from the ground truth, and finally (4) the *user* mode uses real relevance feedback from users. The selection is done on the first 20 videos, which is around the number of videos a user would initially consider.

For the *User* mode, the task of a group of participants was to select relevant and non-relevant videos. 24 results were shown initially, and more could automatically be loaded by scrolling to the bottom of the page. Ten male participants (age $= 26.3, \sigma = 1.567$) with mainly Dutch origin and at least a Bachelor's degree or higher without dyslexia, colour-blindness, concentration problems, or RSI problems, voluntarily participated in an experiment. The participants had two conditions, which correspond to the re-ranking results by ARF and RS. In each of the conditions, 16 queries, randomly assigned using a Latin rectangle [1], were presented to the user, after which they performed relevance feedback.

Evaluation. To evaluate our algorithm, performance of the algorithms is measured by the following aspects: (1) accuracy; (2) robustness. For accuracy, we use Mean Average Precision (MAP). MAP is the standard evaluation method in TRECVID MED and is based on the rank of the positive videos. With re-ranking, the videos that are indicated as positive are always on the top of the list, increase MAP. It is, however, also interesting to know whether the algorithm is able to retrieve new relevant videos. This is why we introduce a variant, *MAP**. MAP* calculates MAP disregarding the videos that have been viewed by the user already, which we track in our experiment. We evaluate our method on both metrics.

For robustness, we report the robustness index (RI) [16]:

$$RI = \frac{|Z_P| - |Z_N|}{|Z|}, \tag{4}$$

where Z_P and Z_N are the sets of queries where the performance difference between ARF and RS in terms of MAP was positive for ARF or negative, respectively, and $|Z|$ is the total number of queries.

3.3 Results

Accuracy. Table 1 shows the MAP^*, and MAP for the relevance feedback algorithms in the *Optimal* relevance feedback mode. A Shapiro-Wilk test was used to assess normality of our precision scores, The assumption of normality is

Fig. 1. Relevance feedback and results for *Working on a woodworking project.*

violated in MEDTRAIN for Initial for MAP ($p < 0.0005$), and for all modes for MAP^* ($p < 0.0005; p = 0.01; p = 0.001$, for Initial, ARF and RS respectively) and for all modes in MEDTEST set ($p < 0.0005$ for MAP and MAP^*). On both the MEDTRAIN and MEDTEST set, the non-parametric Friedman test showed a significant difference for MAP^* ($\chi^2 = 13.528, p = 0.001$ in MEDTRAIN; $\chi^2 = 13.241, p = 0.001$ in MEDTEST) and MAP ($\chi^2 = 18.063, p < 0.0005$ in MEDTRAIN; $\chi^2 = 18.000, p < 0.0005$ in MEDTEST). A Wilcoxon Signed-Ranks Test with Bonferroni correction (0.05/3) showed a significant difference between **ARF and Initial** in MEDTRAIN ($Z = -4.135, p < 0.0005$ for $MAP^*; Z = -4.252, p < 0.0005$ for MAP), and between **ARF and RS** in both MEDTRAIN and MEDTEST($Z = -2.450, p = 0.014$ for $MAP^*; Z = -3.123, p = 0.002$ for MAP in MEDTRAIN; $Z = -2.427, p = 0.015$ for $MAP^*; Z = -3.509, p < 0.0005$ for MAP in MEDTEST).

Table 1. %MAP using the *Optimal* mode.

Algorithm	MEDTRAIN		MEDTEST	
	MAP^*	MAP	MAP^*	MAP
Initial	15.24	18.06	3.47	6.28
RS	16.74	20.30	2.49	6.80
ARF	**18.92**	**24.22**	**3.78**	**8.80**

Table 2 shows the MAP^* for the different relevance feedback algorithms for the relevance feedback methods. Significance is in line with MAP type results, except we observe that for MEDTEST with Pseudo-relevance selection, both relevance feedback methods do not improve on Initial ranking.

Table 2. $\%MAP^*$ scores for Initial, RS and ARF.

Algorithm	MEDTRAIN			MEDTEST		
	Optimal	Pseudo	Random	Optimal	Pseudo	Random
Initial	15.24	15.69	15.69	3.47	**3.42**	3.42
RS	16.74	14.18	14.35	2.49	2.73	3.15
ARF	**18.92**	**18.11**	**18.15**	**3.78**	3.17	**4.36**

Table 3. User experiment $\%MAP^*$ scores and standard deviations.

Algorithm	MAP^*	σ
Initial	13.09	1.02
RS	10.71	1.98
ARF	**15.32**	**1.55**

Table 3 shows the scores found in the user experiments. A Shapiro-Wilk test showed that the precision score distributions do not deviate significantly from a normal distribution at $p > 0.05$ ($p = 0.813; p = 0.947; p = 0.381$, for Initial, RS, and ARF respectively). A statistically significant difference between groups was determined by a one-way ANOVA ($F(2,27) = 18.972$, $p < 0.0005$). A post-hoc Tukey's HSD test was performed to verify intergroup differences. The means of all algorithms differed significantly at $p < 0.05$ ($p = 0.006; p = 0.01; p < 0.0005$, for Initial-RS, Initial-ARF, and RS-ARF, respectively).

In these user experiments, on average, 61.65% of marked relevant results were correct, and 92.71% of marked non-relevant results were correct. Further investigation was done to research the effect of the positive and negative annotations on precision scores. Variations of the α and β parameters were analyzed. While performance decreased slightly when disregarding all positive annotations ($\alpha = 0.0$), it dropped drastically when disregarding the negative annotations ($\beta = 0.0$). In line with relevant literature on Rocchio, $\alpha = 1.15$ and $\beta = 0.5$ provides the highest MAP^*. Visualizations of these results can be found in Fig. 2.

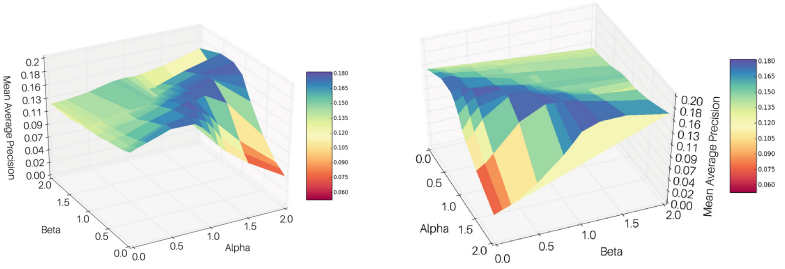

Fig. 2. MAP^* relative to α and β values.

Robustness. To get an overview of the precision per event, we calculated precision averaging over all sessions. A bar plot of RS scores subtracted from ARF scores is shown in Fig. 3. RI was calculated with respect to Initial ranking for ARF and RS, respectively. We see that ARF improves Initial ranking in 71.88% of events, and RS in 37.5%. These scores result in a robustness performance of $RI = 0.4375$ for ARF and $RI = -0.25$ for RS.

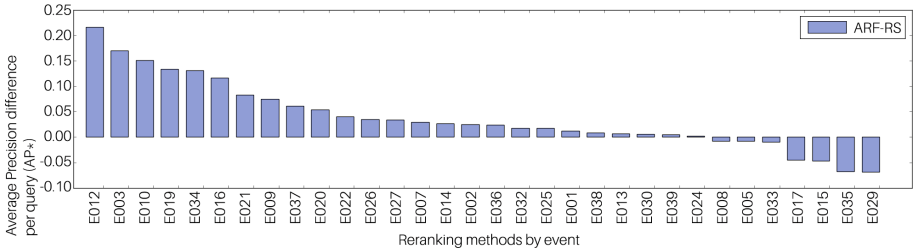

Fig. 3. Average precision difference (AP^*) per event.

Run Time. Since the run-time of the ARF method depends mainly on the size of the video collection ($O(n)$, concept detector weights are updated only based on the selected (non-)relevant results), it is quicker than the RS method whose run-time depends on both total video set size and concept detector set size ($O(n * d)$). Note that the similarity measure applied for RS also factors into this. The average run time for RS and ARF (on an Intel Core i7-4700MQ CPU @ 2.40 GHz x-64 system with 8 GM RAM) is 8003.45 ms and 107.25 ms, respectively.

3.4 Subjective User Evaluation

In additional subjective user evaluation experiments, we also compare ARF results to RS results directly. We asked users (N = 19) to perform relevance selection for as much events as they would like on the initial result set and showed both ARF and RS reranking results. The order in which the events, and reranking results (left or right) were shown was randomized. We then asked users to select the ranking that they thought was the best. Figure 1 shows an example of retrieval results for Initial, RS reranking, and ARF reranking.

We found that ARF is selected 85.96%, of which 83.67% was correct when compared to actual AP scores. RS was preferred in only 14.04% of all queries, of which 25.0% was correct. We also investigated relevance selection per event. An overview of the True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) scores are shown in Fig. 4. Note that this terminology

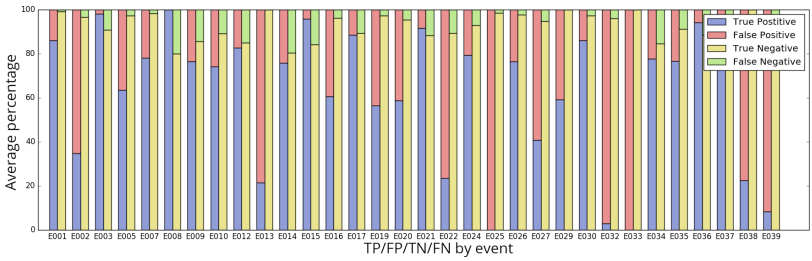

Fig. 4. Percentage relevance selection per event per method.

does not capture the users' beliefs sufficiently, since the user's cannot be wrong in their judgment (if we assume they performed the task honestly). We can state that they are True, or False Positives only relative to the ground truth. A better terminology reflecting the users' honest evaluation might be: *Correct positives*, *Missed negatives*, *Correct negatives* and *Missed positives*.

4 Discussion

Our results show a statistically significant difference ($p < 0.05$) in the means of the MAP^* scores of the ARF algorithm, and that of the Initial and RS algorithm in the user experiments with a relatively small sample size. These results are encouraging, and provide a solid basis for the claim that the ARF algorithm has on average a better performance compared to the RS algorithm. We show that even when discarding events from the MEDTEST set that have a very high accuracy (because we have a high-matching concept detector), we still obtain very reasonable MAP scores. This claim is strengthened by the performance on MAP^*, and MAP in results from experiments with different relevance feedback modes. We see a similar trend between different modes of relevance feedback. Using the *Optimal* relevance feedback mode, ARF performs better than RS and Initial ranking, but not by much. However, when we introduce non-optimal, and perhaps more realistic, relevance feedback modes such as pseudo- or random-relevance feedback, we see ARF performing significantly better. We believe that this effect could be explained by the ARF algorithm being less volatile, and less subject to the effect of misclassification of a single result rippling through to ranking scores. This effect can also be seen in our user experiments, since user relevance feedback is not error-free by a long measure (see Sect. 3.3). In our experiments where we directly let users choose between two reranked sets, they select ARF as the best ranking 85.96% of the time. On the MEDTEST set, we can see comparable results, except that on the *Pseudo* relevance feedback mode, ARF cannot improve on the Initial ranking due to the small number of positive videos at the top of the initially returned results.

Individual events analysis was also performed. There is a clear tendency of the MAP^* to decline when the True Positive (TP) rate drops. While the ARF algorithm obtains the highest MAP^* on average, we see some events on which the RS algorithm, or even Initial ranking performs better. For example, on event $E035$, *Horse riding competition*, RS outperforms ARF. Manual inspection of the initial video set shows that for these events, relevant and non-relevant videos in the initial set are quite homogeneous regarding concept detector scores, but easily distinguishable by human observers. These findings indicate the absence of good concept detectors that capture this distinction.

Dalton et al. [3] show that pseudo-relevance feedback can increase MAP up to 25%, and up to 55% for real relevance feedback. Though we do not match these numbers on average, for some events we gain a performance boost of up to 54% with pseudo- and up to 25% with real-relevance feedback. One striking observation we make is that although MAP(*) scores show a significant difference

between the ARF and RS methods, subjective user evaluation when integrated in a complete video search system does not reflect these differences. However, when asked to compare reranking results directly side-by-side, we do see a preference.

Parameter tuning of the α and β parameters shows the importance of including the non-relevant class in ARF. We see a large decline in performance when disregarding the negative annotations, while this decline was considerably less steep when disregarding the positive annotations (see Sect. 3.3).

5 Conclusion and Future Work

This paper investigated relevance feedback algorithms for video event retrieval. We proposed the adaptive relevance feedback (ARF) algorithm derived from the well-researched Rocchio text retrieval algorithm [15]. While state-of-the-art algorithms in video relevance feedback use few examples from nearest neighbours, ARF updates ranking functions based on the difference between the centroids of the relevant and non-relevant examples.

We investigated several feedback modes, including feedback from real users, on the training set of the Multimedia Event Detection task [13]. We compared ARF against the state-of-the-art algorithms from Gia et al. [5,6], referred to as RS. On the MEDTRAIN and MEDTEST sets, ARF showed stronger average effectiveness compared to RS in terms of MAP^*, and MAP, for a number of different modes of relevance feedback. This effect was also found in subjective user evaluation experiments. Robustness for ARF was also higher compared to RS.

From the above experiments we conclude that there is strong evidence that ARF shows better performance for video event retrieval than the previous state-of-the-art. The performance trends between real and pseudo relevance feedback modes were similar for both tested algorithms, which is a secondary contribution of this paper.

For future work, we propose to evaluate ARF on a broader range of datasets in terms of videos and concept detectors. Furthermore, discovering concept detectors with poor performance through relevance feedback, and adding supplementary concept detectors outside the initial set should be researched. Another interesting avenue for further research is the apparent contrast between objective MAP scores and subjective user evaluation. Do users 'care' about MAP scores?

References

1. Cochran, W.G., Cox, G.M.: Experimental designs (1957)
2. Crucianu, M., Ferecatu, M., Boujemaa, N.: Relevance feedback for image retrieval: a short survey. Report of the DELOS2 European Network of Excellence (FP6) (2004)
3. Dalton, J., Allan, J., Mirajkar, P.: Zero-shot video retrieval using content and concepts. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, pp. 1857–1860. ACM (2013)

4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR 2009, pp. 248–255. IEEE (2009)
5. Deselaers, T., Paredes, R., Vidal, E., Ney, H.: Learning weighted distances for relevance feedback in image retrieval. In: 19th International Conference on Pattern Recognition, ICPR 2008, pp. 1–4. IEEE (2008)
6. Gia, G., Roli, F., et al.: Instance-based relevance feedback for image retrieval. In: Advances in Neural Information Processing Systems, pp. 489–496 (2004)
7. Jiang, L., Meng, D., Mitamura, T., Hauptmann, A.G.: Easy samples first: self-paced reranking for zero-example multimedia search. In: Proceedings of the ACM International Conference on Multimedia, pp. 547–556. ACM (2014)
8. Jiang, L., Mitamura, T., Yu, S.I., Hauptmann, A.G.: Zero-example event search using multimodal pseudo relevance feedback. In: Proceedings of the International Conference on Multimedia Retrieval, p. 297. ACM (2014)
9. Jiang, L., Yu, S.I., Meng, D., Mitamura, T., Hauptmann, A.G.: Bridging the ultimate semantic gap: a semantic search engine for internet videos. In: ACM International Conference on Multimedia Retrieval, pp. 27–34 (2015)
10. Jiang, Y.G., Wu, Z., Wang, J., Xue, X., Chang, S.F.: Exploiting feature and class relationships in video categorization with regularized deep neural networks. arXiv preprint arXiv:1502.07209 (2015)
11. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014)
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
13. Over, P., Awad, G., Michel, M., Fiscus, J., Sanders, G., Kraaij, W., Smeaton, A.F., Quéenot, G., Ordelman, R.: TRECVID 2015 - an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of the TRECVID 2015, p. 52. NIST, USA (2015)
14. Patil, S.: A comprehensive review of recent relevance feedback techniques in CBIR. Int. J. Eng. Res. Technol. (IJERT) **1**(6) (2012)
15. Rocchio, J.J.: Relevance feedback in information retrieval (1971)
16. Sakai, T., Manabe, T., Koyama, M.: Flexible pseudo-relevance feedback via selective sampling. ACM Trans. Asian Lang. Inf. Process. (TALIP) **4**(2), 111–135 (2005)
17. Tao, D., Tang, X., Li, X.: Which components are important for interactive image searching? IEEE Trans. Circuits Syst. Video Technol. **18**(1), 3–11 (2008)
18. Tong, S., Chang, E.: Support vector machine active learning for image retrieval. In: Proceedings of the 9th ACM International Conference on Multimedia, pp. 107–118. ACM (2001)
19. Wang, X.Y., Liang, L.L., Li, W.Y., Li, D.M., Yang, H.Y.: A new SVM-based relevance feedback image retrieval using probabilistic feature and weighted kernel function. J. Vis. Commun. Image Represent. **38**, 256–275 (2016)
20. Xu, S., Li, H., Chang, X., Yu, S.I., Du, X., Li, X., Jiang, L., Mao, Z., Lan, Z., Burger, S., et al.: Incremental multimodal query construction for video search. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 675–678. ACM (2015)
21. Yang, L., Hanjalic, A.: Supervised reranking for web image search. In: Proceedings of the International Conference on Multimedia, pp. 183–192. ACM (2010)
22. Ye, G., Liu, D., Chang, S.F., Saleemi, I., Shah, M., Ng, Y., White, B., Davis, L., Gupta, A., Haritaoglu, I.: BBN VISER TRECVID 2012 multimedia event detection and multimedia event recounting systems

23. Zhang, H., Lu, Y.J., de Boer, M., ter Haar, F., Qiu, Z., Schutte, K., Kraaij, W., Ngo, C.W.: VIREO-TNO@ TRECVID 2015: multimedia event detection
24. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems, pp. 487–495 (2014)
25. Zhou, X.S., Huang, T.S.: Relevance feedback in image retrieval: a comprehensive review. Multimed. Syst. **8**(6), 536–544 (2003)