18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

Adaptive Relevance Feedback for Fusion of Text and Visual Features

Leszek Kaliciak[‡*], Hans Myrhaug[‡], Ayse Goker[‡*], and Dawei Song[§¶]
[*] The Robert Gordon University, Aberdeen, Scotland
Email: l.kaliciak@rgu.ac.uk; a.s.goker@rgu.ac.uk
[‡]Ambiesense Limited, Aberdeen, Scotland
Email: hans@ambiesense.com
[§]Tianjin University, Tianjin, China
[¶]The Open University, Milton Keynes, UK
Email: dawei.song@open.ac.uk

***Abstract*—It has been shown that query can be correlated with its context to a different extent; in this case the feedback images. We introduce an adaptive weighting scheme where the respective weights are automatically modified, depending on the relationship strength between visual query and its visual context and textual query and its textual context; the number of terms or visual terms (mid-level visual features) co-occurring between current query and its context. The user simulation experiment has shown that this kind of adaptation can indeed further improve the effectiveness of hybrid CBIR models.**

***Keywords*—Hybrid Relevance Feedback, Visual Features, Textual Features, Early Fusion, Late Fusion, Re-Ranking, Adaptive Weighting Scheme**

I. Introduction

The semantic gap in Content-based Image Retrieval (CBIR), the difference between human perception and machine representation of multimedia objects, can be reduced by intelligently combining low level visual features and high level semantic information - textual features. Visual and textual feature spaces are complementary and correlated, therefore an ideal combination should exploit these relationships to further improve CBIR performance.

We can further reduce the aforementioned semantic gap by letting the user interact with the retrieval system. Collected user implicit or explicit feedback can be then utilized to narrow down the search and contextualize results according to user's interests and preferences. In this paper, the visual and textual context is generated by the visual and textual representations of feedback images, respectively. Moreover, we can combine the visual and textual features in the context of relevance feedback. Apart from the visual and textual query representations, we would now obtain visual and textual context subspaces of feedback images. This makes it possible to exploit inter (visual-textual) and intra (visual-visual, textual-textual) correlations between these subspaces.

A hybrid model for the combination of visual and textual features in the context of user feedback that exploits the aforementioned correlations, has been introduced by Kaliciak et al. [10]. It was proven effective as it outperformed other hybrid models which could be modified to incorporate user feedback. We will be referring to this model as hybrid CBIR relevance feedback model. Although our experiments are based on explicit user feedback, the same hybrid model can also utilize query history (implicit user feedback). In this case, the contextual feature subspaces would consist of visual and textual image representations from the query history.

This paper is a follow-up on our previous work, an enhancement of the fixed weights hybrid relevance feedback model.

The hybrid CBIR relevance feedback model for the visual and textual features combination utilizes fixed weights corresponding to the importance of query and its context (here, feedback images). It has been shown, however, that query can be correlated with its context to a different extent ([17], [7], [6], [5]). Goker has stated the importance of (query) context in meeting users' information needs and how queries do not occur in isolation. This can be exploited to improve text retrieval.

In general, the importance of the original query and its context should not be fixed for all the queries. Inspired by this observation, we aim to develop an adaptive weighting scheme to further improve the hybrid CBIR relevance feedback model. Thus, each query would be associated with a unique set of weights corresponding to the relationship strength between visual query and its visual context as well as the textual query and its textual context. The higher the number of textual or visual terms that co-occur between current query and the context, the stronger the relationship and vice versa. If the relationship between query and its context is weak, context becomes important. We then adjust the probability of the original query terms, and the adjustment will significantly modify the original query. However, if the aforementioned relationship (similarity) between query and its context is strong, context will not help much. The original query terms will tend to dominate the whole term distribution in the modified model. The adjustment will not significantly modify the original query. In this paper, the textual and visual terms refer to image tags and instances of visual words, respectively. Visual words are related to the "bag of visual words" framework and do not possess semantic meaning. They are the most representative local visual patterns in the image collection.

We tested the enhanced model with adaptive weighting scheme within a user feedback simulation framework. For fair comparison purposes, the best performing sets of fixed weights were selected for evaluation against the new model. We have

shown that our enhanced model can outperform the original one with fixed weights.

Our contribution in this paper is related to describing how to measure the relationship strength between query and its context, and how to incorporate the adaptive weighting scheme into the state-of-the-art existing model to further improve the retrieval.

The rest of this paper is organized as follows. Section 2 presents related work. Section 3 introduces the original model for the combination of features in the context of user feedback and presents the enhanced model with adaptive weighting scheme. Section 4 describes the experimental setup and results obtained on the ImageCLEF data collection with their analysis. Conclusions are drawn in Section 5, along with the future work (Section 6).

II. Related Work

Most approaches in CBIR that utilize user feedback are mono-modal, as they exploit only visual or only textual feedback representations. However, it is possible to modify some existing fusion methods in order to incorporate user feedback and combine the features in the user feedback context. This was done in [10], where the original hybrid model with fixed weights was compared against other hybrid models which incorporate user feedback, and modified hybrid models. Let us briefly introduce the widely used feature combination methods.

Pre-filtering by text and re-ranking by visual content is usually a well performing method. However, the main drawback of this approach is that the images without the textual description will never be returned by the system although one could try to auto-annotate the collection beforehand. Moreover, this type of pre-filtering relies heavily on the textual features and the assumption that images are correctly annotated.

The most common early fusion technique is concatenation of visual and textual representations. Some recently proposed models incorporate the tensor product to combine the systems [18]. Tensor product captures the relationships between all dimensions of different feature spaces. The main drawback of the early fusion approach, however, is the well-known curse of dimensionality.

In the case of late fusion, the most widely used method is the arithmetic mean of the scores, their sum (referred to as CombSUM), or their weighted linear combination. One of the best performing systems on the ImageCLEF data collection, XRCE [12], utilizes both (for comparison purposes) early (concatenation of features) and late (an average of scores) fusion approaches. Another common combination method, referred to as CombPROD in the literature, is the square of the geometric mean of the scores - their product. It has been argued that the major drawback of the late fusion approaches is their inability to capture the correlation between different modalities [13].

It has been discovered, however, that specific early and late fusion strategies can be interchangeable [11].

Other combination methods involve a combination of late fusion and image re-ranking [3]. Because the first stage is based on the pre-filtering of the collection by text, the model is referred to as the semantic combination.

The feature combination methods can be also considered in the context of classification, e.g. image categorization [1]. In the case of classification, late fusion (for example) of features is performed differently, as a weighted voting strategy from the outputs of different classiers ([15], [16]). Some fusion strategies in CBIR can be also classified as intermediate fusion [1]. They simultaneously learn individual classier and combination classier weights [21], and this process happen at various levels of learning. In this paper, however, our focus is on the similarity-based image retrieval, not the classification.

The fusion approach that can be easily modified to incorporate the user feedback is based on the so-called transmedia pseudo-relevance mechanism. This feedback query expansion is based on textual query expansion in most of the papers ([4],[2]). Typically, textual annotations from the top visually-ranked images (or from a mixed run) are used to expand a textual query.

Similarly to research involving models which combine the features in the context of user feedback, there is not much research on the adaptive combination of query and its context. We are going to mention a few mono-modal approaches that utilize adaptive weighting schemes.

Wu et al. [20] implement an adaptive data fusion method with dynamically adjustable weights. They investigate two methods for the weight updating, namely "performance square" updating and a mixture of the aforementioned and linear regression analysis. Experiments conducted on the benchmark showed that both adaptive weights models outperformed CombSUM fusion method. They combine evidence from different sources but do not incorporate any user feedback.

Wang et al. [19] proposed an adaptive weighting approach to improve the current statistical context-sensitive retrieval model. They first investigate the so-called "potential for adaptability", the performance gap between the context-sensitive model with fixed weights and the one with adaptive weights, to show that the system can really benefit from having query-specific weights. They apply the support vector regression to build a weight-prediction model, which enables a more flexible combination of current query and its context.

Most approaches that try to adapt the weights corresponding to query and its context have the linear combination of the relevance scores at their core. Machine learning is then often used to dynamically change these weights.

Our adaptive weighting approach differs from the above in that it represents a hybrid approach; it is not mono-modal. Moreover, we utilize the state-of-the-art hybrid approach that takes into account the inter- and intra-correlations between feature spaces and combine them in the context of user feedback which is different from simply combining them in an ad-hoc manner. We utilize two notions of user feedback, visual and textual.

III. HYBRID RELEVANCE FEEDBACK WITH ADAPTIVE WEIGHTING SCHEME

The hybrid CBIR relevance feedback model proved to be an effective tool in the semantic gap reduction [10]. In this section, we briefly describe the model for the combination of features in the context of user feedback. Then, we show how

to enhance that approach by incorporating adaptive weighting scheme. Instead of having arbitrary fixed weights, we can automatically adjust them with respect to the relationship strength between query and its context. This will allow us to further improve the retrieval's effectiveness.

A. Hybrid Relevance Feedback

The original hybrid relevance feedback model is defined on a Hilbert space (Hilbert space is usually defined as a complex space with an inner product) which can be thought of as a natural extension of the standard vector space model, with its useful notions of subspaces and projections. It was inspired by the mathematical tools utilized in Quantum Mechanics (QM) and is based on the expectation value, predicted mean value of the measurement. The model is based on the notion of co-occurrence and the tensor operation. Co-occurrence matrices can be treated as density matrices (probability distribution) because they are Hermitian and positive-definite, and the tensor operator $\otimes$ can be utilized to combine the density matrices corresponding to visual and textual feature spaces. In quantum mechanics, the tensor product of density matrices of different systems represents a density matrix of the combined system.

Thus, the intra-feature correlations are captured by density matrices corresponding to individual feature spaces, and inter-correlations are modeled in the form of the tensor product - resulting in a density matrix of the composite system. The projection of a query onto the subspace of the composite system can then be considered as our similarity measurement.

Thus, in the fixed weights model, the similarity measurement on the combined space is given by

$$tr\left((\otimes_n M_n)\cdot\left(\otimes_n\left(a_n^T\cdot a_n\right)\right)\right)=\prod_n\left\langle M_n\middle|a_n^T\cdot a_n\right\rangle \tag{1}$$

where tr denotes the matrix trace operator, M_n for $n = 2,3,\ldots,N_0; N_0 \in N$ are defined as weighted combinations of co-occurrence matrices corresponding to different feature spaces (a subspace generated by the query vector and vectors from the feedback set), $\otimes$ denotes the tensor operator, and a_n represent particular image feature.

Now, let q_v, q_t denote the visual and textual representations of the query; c^i, d^i denote visual and textual representations of the images in the feedback set; D_q^v, D_f^v denote the density (co-occurrence) matrices of a visual query and its visual context (feedback images); D_q^t, D_f^t denote the density matrices of a textual query and its textual context; r_1, $1-r_1$ (r_2, $1-r_2$) denote the weighting factors (constant, importance of query and feedback density matrices respectively); and n denote the number of images in the feedback set. Then, M_1 and M_2 can be defined as

$$M_1 = r_1\cdot D_q^v+\frac{1-r_1}{n}\cdot D_f^v= r_1\cdot q_v^T\cdot q_v+\sum_i\left(\frac{1-r_1}{n}\cdot\left(c^i\right)^T\cdot c^i\right) \tag{2}$$

and

$$M_2 = r_2\cdot D_q^t+\frac{1-r_2}{n}\cdot D_f^t= r_2\cdot q_t^T\cdot q_t+\sum_i\left(\frac{1-r_2}{n}\cdot\left(d^i\right)^T\cdot d^i\right) \tag{3}$$

The model can be represented as

$$\left\langle M_1\otimes M_2\middle|\left(a^T\cdot a\right)\otimes\left(b^T\cdot b\right)\right\rangle= \left(r_1\cdot\langle q_v|a\rangle^2+\frac{1-r_1}{n}\cdot\sum_i\left\langle c^i\middle|a\right\rangle^2\right)\cdot \left(r_2\cdot\langle q_t|b\rangle^2+\frac{1-r_2}{n}\cdot\sum_i\left\langle d^i\middle|b\right\rangle^2\right) \tag{4}$$

Hence, this hybrid relevance feedback breaks down into the weighted combinations of measurements performed on individual feature spaces. The squared inner products can be interpreted as the probabilities of the system transitions from one state to another (here, similarity measures).

B. On the Importance of Query and Its Context. Adaptive Weighting Scheme

In the original model, the weights corresponding to textual query, textual context, visual query and visual context are fixed (i.e. r_1, $1-r_1$, r_2, $1-r_2$, across all the queries). However, it has been highlighted [17], that the query may be correlated with its context to a different extent. In this paper, terms "textual query", "visual context" etc. refer to the textual (visual) representation of a query image (context images).

We can further improve the feedback model by adjusting these weights with respect to the issued queries and feedback images based on the strength of the relationship between query and its context.

We will measure the strength of the aforementioned relationship by computing the similarity between co-occurrence matrices corresponding to the query and its context (feedback images). The higher the number of terms or visual terms (mid-level features) co-occurring between current query and the context, the stronger the relationship and vice versa. Thus, the relationship strength between the visual query and visual context can be measured as

$$\left\langle D_q^v\middle|D_f^v\right\rangle=\left\langle q_v^T\cdot q_v\middle|\sum_i\left(c^i\right)^T\cdot c^i\right\rangle= \sum_i\left\langle q_v\otimes q_v\middle|c^i\otimes c^i\right\rangle= \sum_i\left\langle q_v\middle|c^i\right\rangle^2 \tag{5}$$

Similarly, the relationship strength between the textual query and its textual context can be computed as

$$\left\langle D_q^t\middle|D_f^t\right\rangle=\sum_i\left\langle q_t\middle|d^i\right\rangle^2 \tag{6}$$

We can normalize these measurements. Thus, we can compute

$$\frac{\left\langle D_q^v \middle| D_f^v \right\rangle}{\left\| D_q^v \right\| \cdot \left\| D_f^v \right\|} \tag{7}$$

$$\frac{\left\langle D_q^t \middle| D_f^t \right\rangle}{\left\| D_q^t \right\| \cdot \left\| D_f^t \right\|} \tag{8}$$

where

$$\left\| D_q^v \right\| = \sqrt{\left\langle D_q^v \middle| D_q^v \right\rangle} = \sqrt{\langle q_v^T \cdot q_v | q_v^T \cdot q_v \rangle} = \sqrt{\langle q_v \otimes q_v | q_v \otimes q_v \rangle} = \sqrt{\langle q_v | q_v \rangle^2} = \langle q_v | q_v \rangle \tag{9}$$

and

$$\left\| D_f^v \right\| = \sqrt{\left\langle D_f^v \middle| D_f^v \right\rangle} = \sqrt{\left\langle \sum_i \left(c^i\right)^T \cdot c^i \middle| \sum_i \left(c^i\right)^T \cdot c^i \right\rangle} = \sqrt{\sum_i \sum_i \langle c^i \otimes c^i | c^i \otimes c^i \rangle} = \sqrt{n \cdot \sum_i \langle c^i | c^i \rangle^2} \tag{10}$$

Analogically, for the textual part

$$\left\| D_q^t \right\| = \langle q_t | q_t \rangle \tag{11}$$

and

$$\left\| D_f^t \right\| = \sqrt{n \cdot \sum_i \langle d^i | d^i \rangle^2} \tag{12}$$

Thus, the modified model becomes

$$tr\left(\left(M_1 \otimes M_2\right)\left(\left(a^T a\right) \otimes \left(b^T b\right)\right)\right) = \left(str_v \langle q_v | a \rangle^2 + \left(1 - str_v\right) \frac{1}{n} \sum_i \left\langle c^i \middle| a \right\rangle^2\right) \cdot \left(str_t \langle q_t | b \rangle^2 + \left(1 - str_t\right) \frac{1}{n} \sum_i \left\langle d^i \middle| b \right\rangle^2\right) \tag{13}$$

where

$$str_v = \frac{\left\langle D_q^v \middle| D_f^v \right\rangle}{\left\| D_q^v \right\| \left\| D_f^v \right\|} = \frac{\sum_i \left\langle q_v \middle| c^i \right\rangle^2}{\langle q_v | q_v \rangle \sqrt{n \sum_i \langle c^i | c^i \rangle^2}} \tag{14}$$

and

$$str_t = \frac{\left\langle D_q^t \middle| D_f^t \right\rangle}{\left\| D_q^t \right\| \left\| D_f^t \right\|} = \frac{\sum_i \left\langle q_t \middle| d^i \right\rangle^2}{\langle q_t | q_t \rangle \sqrt{n \sum_i \langle d^i | d^i \rangle^2}} \tag{15}$$

Let us assume that the relevance feedback is given after the first round retrieval to refine the query. The adaptive weighting can be interpreted in a following way:

1) small $\langle D_q | D_f \rangle$; weak relationship between query and its context, context becomes important. We adjust the probability of the original query terms; the adjustment will significantly modify the original query.
2) big $\langle D_q | D_f \rangle$; strong relationship (similarity) between query and its context, context will not help much. The original query terms will tend to dominate the whole term distribution in the modified model. The adjustment will not significantly modify the original query.

The experiments conducted on ImageClef data collection have shown that the adaptive weighting can indeed outperform the fixed weighting scheme (in the context of hybrid models and user feedback).

C. Generalization of the Model

Our enhanced model can be naturally expanded to accommodate other features, i.e. various visual features

$$tr\left(\left(\otimes_n M_n\right) \cdot \left(\otimes_n \left(a_n^T \cdot a_n\right)\right)\right) = \prod_n \langle M_n | a_n^T \cdot a_n \rangle \tag{16}$$

Thus, for 3 features (i.e. two visual and a textual feature) our enhanced model becomes

$$tr\left(\left(M_1 \otimes M_2 \otimes M_3\right)\left(\left(a_1^T a_1\right) \otimes \left(a_2^T a_2\right) \otimes \left(b^T b\right)\right)\right) = \left(str_{v1} \langle q_{v1} | a_1 \rangle^2 + \left(1 - str_{v1}\right) \frac{1}{n} \sum_i \left\langle c_1^i \middle| a_1 \right\rangle^2\right) \cdot \left(str_{v2} \langle q_{v2} | a_2 \rangle^2 + \left(1 - str_{v2}\right) \frac{1}{n} \sum_i \left\langle c_2^i \middle| a_2 \right\rangle^2\right) \cdot \left(str_t \langle q_t | b \rangle^2 + \left(1 - str_t\right) \frac{1}{n} \sum_i \left\langle d^i \middle| b \right\rangle^2\right) \tag{17}$$

where

$$str_{v1} = \frac{\left\langle D_q^{v1} \middle| D_f^{v1} \right\rangle}{\left\| D_q^{v1} \right\| \left\| D_f^{v1} \right\|} = \frac{\sum_i \left\langle q_{v1} \middle| c_1^i \right\rangle^2}{\langle q_{v1} | q_{v1} \rangle \sqrt{n \sum_i \left\langle c_1^i \middle| c_1^i \right\rangle^2}} \tag{18}$$

and

$$str_{v2} = \frac{\left\langle D_q^{v2} \middle| D_f^{v2} \right\rangle}{\left\| D_q^{v2} \right\| \left\| D_f^{v2} \right\|} = \frac{\sum_i \left\langle q_{v2} \middle| c_2^i \right\rangle^2}{\langle q_{v2} | q_{v2} \rangle \sqrt{n \sum_i \left\langle c_2^i \middle| c_2^i \right\rangle^2}} \tag{19}$$

and

$$str_t = \frac{\left\langle D_q^t \middle| D_f^t \right\rangle}{\left\| D_q^t \right\| \left\| D_f^t \right\|} = \frac{\sum_i \left\langle q_t \middle| d^i \right\rangle^2}{\langle q_t | q_t \rangle \sqrt{n \sum_i \langle d^i | d^i \rangle^2}} \tag{20}$$

Here, for example, M_1, a_1 and M_2, a_2 may correspond to different visual features (density matrices and vector representations of images from the data collection), and M_3, b corresponds to a textual feature.

IV. Experiments and Discussions

We have conducted our proof of concept experiments on a data collection comprising 20000 images. This data-set comes with a realistic ground truth data for the evaluation and is fully annotated. Thus, it is very useful for CBIR proof of concept experiments. We have recently transitioned to 1 million image collection, MIR Flickr - 1m, and started conducting real user evaluation. The initial tests look very promising.

Figure 1 on page 5 shows the refinement of the search results based on the hybrid relevance feedback model for 1 million images data collection.

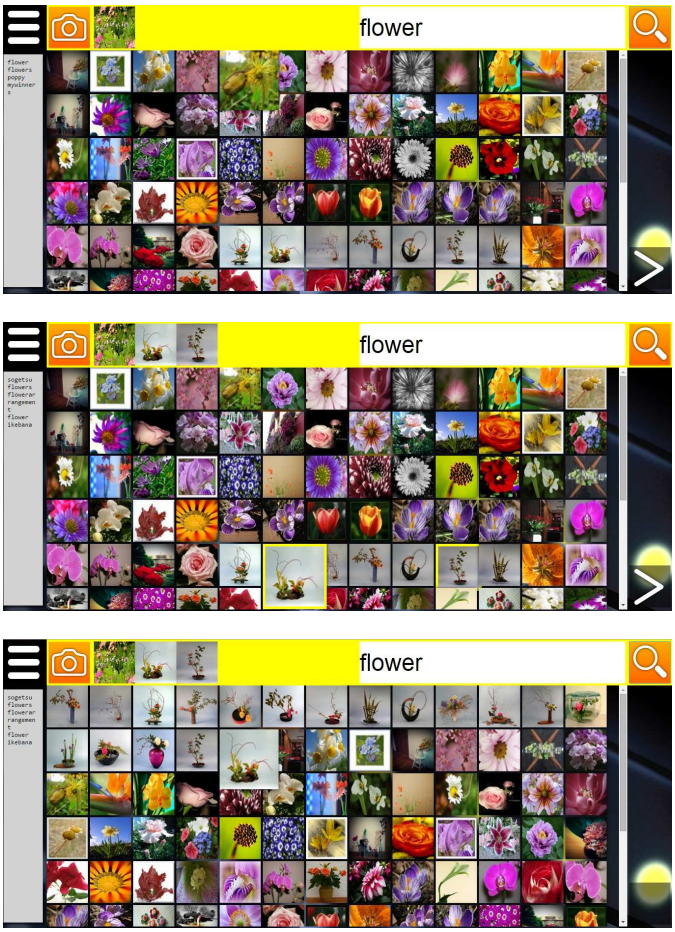

Fig. 1: Hybrid relevance feedback at work, 1 million images. Top: User queries the system by visual example and text. The system retrieves and displays the results. Middle: User selects images for the search refinement. Bottom: Relevant images are "pushed" towards the top-left corner of the results panel (top ranked images). The refinement is based on the combination of visual and textual feature spaces.

A. Data Collection

ImageCLEFphoto2007 consists of 20000 everyday real-world photographs [8]. It is a standard image collection used by Information Retrieval (IR) community for evaluation purposes. There are 60 query topics that do not belong to the collection. Example topics are shown in Table 1. ImageCLEFphoto2007 data collection is considered to be very difficult for retrieval systems because of the abstract semantic content of many queries. For example, the topic "straight road in the USA" could be difficult for visual features whereas "church with more than two towers" could be hard for textual features. This is indeed the motivation why we incorporate hybrid models in CBIR.

TABLE I: Example topics in ImageCLEFphoto2007 data collection

Accommodation with swimming pool
Church with more than two towers
Religious statue in the foreground
People with a flag
Straight road in the USA
Group standing in salt pan
Host family posing for a photo
Tourist accommodation near Lake Titicaca
Destinations in Venezuela
People observing football match

B. Experimental Setup

For the consistency and comparison with the fixed weight model, we test the adaptive weight approach in a simulated user feedback framework. First, we perform the first round retrieval for a topic from the query set based on the visual features only (we retrieve 1000 images). We use the visual features only because in the real life scenario many images would not have textual descriptions. We also do not combine the features in the first round retrieval as this would represent a different task. In this work we want to focus on testing the features' combination models within the user feedback framework.

We identify 1, 2 and 3 relevant images respectively from the highest ranked images based on the ground truth data. Thus obtained images simulate the user feedback and are utilized in the proposed model to re-score the data collection. For each query topic (60 in total) we calculate mean average precision (MAP) for the top 20 retrieved images, as it is unlikely that users would look at more than this number of documents. For the fixed weight model, we test different combinations of parameters and choose the ones performing best for a fair comparison with our adaptive weighting scheme. The MAP is usually considered to be one of the main performance measures of the automatic system evaluations in CBIR. The annual ImageCLEF challenge, for example, ranks the CBIR systems according to the MAP.

The visual features used in the experiment are based on the Bag of Features (BOF) framework (see [9] for a detailed description of the utilized approach). BOF are regarded as mid-level visual features and represent current state-of-the-art in CBIR. The first step in the BOF framework is to localize the so-called points of interest (point-like, region-like) by using corner/blob detectors. Other sampling techniques include random and dense sampling. The second step involves the representation of regions around the sample points in a form of multidimensional vectors by applying certain content descriptors. There are various existing descriptors, the SIFT (Scale Invariant Feature Transform) being one of the widely

used ones. The initial extraction is performed on a training set of images and the K-means clustering is applied to it. Each cluster will correspond to one visual word, a local pattern. Finally, each image in a data collection can be characterized by a histogram of visual words' counts. Here, we utilize random sampling (best in generic image retrieval when the number of sample points is high [14]) to get 900 sample points. We use colour moments as descriptors and generate 40 dimensional vectors of visual words counts.

In addition to the local features, we also experiment with a global method - colour histogram in RGB colour space. First, each image is split into individual colour channels (a grey-scale representation of an individual colour). Next, pixel intensities corresponding to each colour channel are quantized into 8 bins. Thus obtained three histograms are concatenated to form a 24 dimensional colour histogram.

The textual features were obtained by applying the standard Bag of Words technique, with Porter stemming, stop words removal, and term frequency - inverse document frequency weighting scheme.

C. Baselines and Models for Comparison Purposes

Let *prMMFixed* and *prMMAdapt* denote the hybrid CBIR relevance feedback model with fixed weights and the enhanced model with an adaptive weighting scheme, respectively. We also test the adaptive capabilities of the visual and textual elements of the model. Thus, by *vOnlyFix*, *tOnlyFix* we will denote the visual and textual parts of the model with fixed weights and *vOnlyAd*, *tOnlyAd* will represent visual and textual parts of the model with adaptive weighting scheme.

Early fusion is represented by a modified Rocchio algorithm (*eFus*). In the Rocchio algorithm we use the most common weight of the positive context, which is 0.8. The only difference between this variation and the classic model is that it is applied to concatenated visual and textual vectors, as opposed to visual or textual representations only. Let $\oplus$ denote the concatenation operation. Then, this model modify the query in a following way

$$newQuery = q_v \oplus q_t + \frac{0.8}{n} \sum_i (c_i \oplus d_i) \tag{21}$$

After the query modification the scores are recomputed.

Another baseline, which we will refer to as *lFus* will be represented as a combination of all the scores

$$sim\,(q_v, a) + sim\,(q_t, b) + \frac{0.8}{n} \sum_i sim\,(c_i, a) + \frac{0.8}{n} \sum_i sim\,(d_i, b)$$

where sim denotes the similarity between given vectors. In this work sim is an inner product between two vectors.

We can observe that the performance of the two aforementioned baselines must be exactly the same. This stems from the fact, that

$$\begin{aligned} newQuery &= q_v \oplus q_t + \frac{0.8}{n} \sum_i (c_i \oplus d_i) \\ imagesInDataset &= a \oplus b \quad forAll \quad a, b \in Dataset \end{aligned} \tag{22}$$

$$\begin{aligned} &\langle newQuery | imagesInDataset \rangle = \\ &\left\langle q_v \oplus q_t + \frac{0.8}{n} \sum_i (c_i \oplus d_i) \middle| a \oplus b \right\rangle = \\ &\langle q_v \oplus q_t | a \oplus b \rangle + \frac{0.8}{n} \sum_i \langle c_i \oplus d_i | a \oplus b \rangle = \\ &\langle q_v | a \rangle + \frac{0.8}{n} \sum_i \langle c_i | a \rangle + \langle q_t | b \rangle + \frac{0.8}{n} \sum_i \langle d_i | b \rangle \end{aligned} \tag{23}$$

Thus, in our case the early and late fusion strategies (modified Rocchio algorithm operating on concatenated representations and weighted linear combination of scores) are interchangeable.

Our third baseline *rrText* denotes the re-ranking of the results obtained from the first round retrieval based on the aggregated textual representations of the feedback images. Similarly, *rrVis* represents re-ranking of the top retrieved images based on the aggregated visual representations of the images from the feedback set.

Next model *trMed* represents the transmedia feedback query modification. Here, textual annotations from the feedback images (identified by visual features) are used to modify a textual query.

The system performance without simulated feedback will be denoted as *noFback*.

D. Experimental Results and Discussion

In this work, the Mean Average Precision (MAP) is calculated for 20 top images only as this is a more realistic scenario (especially for user simulation/user feedback context). For 1000 top and 3 feedback images, the original system's performance is approximately $MAP \approx 0.206$. If we consider the ImageCLEF2007photo results of other systems (the best models utilize both visual and textual information) which can be found on the ImageCLEF website [22], the hybrid CBIR relevance feedback model places itself among the best performing approaches.

First, let us check the performance of individual components of the original model for different values of parameters r_1, r_2 (fixed weights). We will select the best combination of parameters for a fair comparison with the adaptive weighting scheme. Tables 2 and 3 show the performance of visual and textual components of the hybrid CBIR relevance feedback model for different parameters' values (fixed weights). Last row displays the adaptive capabilities of weights r_1, r_2. Significantly different results (adaptive part against the fixed weights models) are displayed in bold font in Tables 2, 3, 4;$p = 0.05$; paired t-test.

We can observe, that different values of parameter r have different impact on the component's performance. Moreover, the visual part of the model with adaptive weights performed significantly better than the fixed weights part for the higher

TABLE II: Simulated Relevance Feedback, ImageCLEF2007photo results (MAP), Visual part only

	1 FbackImg	2 FbackImg	3 FbackImg
$noFback$	0.013	0.013	0.013
$r_1 = 0$	0.041	0.053	0.061
$r_1 = 0.2$	0.036	0.060	0.070
$r_1 = 0.4$	0.036	0.046	0.047
$r_1 = 0.5$	0.033	0.038	0.044
$r_1 = 0.6$	0.027	0.034	0.041
$r_1 = 0.8$	0.020	0.031	0.039
$r_1 = 1$	0.018	0.029	0.038
$\mathbf{r_1 adapt}$	0.036	0.063	**0.081**

TABLE III: Simulated Relevance Feedback, ImageCLEF2007photo results (MAP), Textual part only

	1 FbackImg	2 FbackImg	3 FbackImg
$noFback$	0.013	0.013	0.013
$r_2 = 0$	0.058	0.072	0.075
$r_2 = 0.2$	0.063	0.076	0.079
$r_2 = 0.4$	0.063	0.076	0.081
$r_2 = 0.5$	0.062	0.080	0.080
$r_2 = 0.6$	0.062	0.079	0.080
$r_2 = 0.8$	0.062	0.079	0.080
$r_2 = 1$	0.052	0.071	0.071
$\mathbf{r_2 adapt}$	**0.085**	**0.095**	**0.112**

number of feedback images. The textual part with adaptive weights shows even better adaptive capabilities of its weights.

Table 4 shows the performance of the hybrid CBIR relevance feedback model for different combinations of parameters values (fixed weights), and the results of the enhanced model with adaptive weighting scheme.

TABLE IV: Simulated Relevance Feedback, ImageCLEF2007photo results (MAP)

	1 FbackImg	2 FbackImg	3 FbackImg
$noFback$	0.013	0.013	0.013
$r_1 = 0.2; r_2 = 0.4$	0.080	0.098	0.115
$r_1 = 0.4; r_2 = 0.2$	0.082	0.101	0.114
$r_1 = 0.5; r_2 = 0.5$	0.082	0.097	0.115
$r_1 = 0.2; r_2 = 0.8$	0.081	0.098	0.116
$r_1 = 0.8; r_2 = 0.2$	0.084	0.096	0.113
$r_1 = 0.2; r_2 = 0.2$	0.081	0.096	0.113
$r_1 = 0.8; r_2 = 0.8$	0.084	0.097	0.115
$\mathbf{r_1, r_2 adapt}$	0.091	**0.12**	**0.142**

Although individual components of the hybrid CBIR relevance feedback model exhibited some sensitivity to the changing values of weights r_1, r_2, the combined model's performance is relatively stable, regardless of the values of the fixed weights. However, if the weights are automatically adjusted for each individual query, the enhanced model performs significantly better (for more than two images in the feedback set).

Finally, the overall comparison of different models is shown in Table 5. Note that in Tables 5 and 6, the significantly different results are displayed in bold font ($p = 0.05$; paired t-test). We will denote the statistical significance over the fixed weights model by *. The bold font and the absence of * symbol will then represent the statistical significance over the baselines.

Our main focus here should be on the difference in performance of the original fixed weight model and our enhanced model with adaptive weighting scheme. It has been shown that the hybrid CBIR relevance feedback model outperformed other state-of-the-art hybrid systems that can be modified to incorporate user feedback. Our experiments confirm the previous findings. In general, the enhanced model significantly outperformed the original one (for more than one image in the feedback set).

TABLE V: Simulated Relevance Feedback, ImageCLEF2007photo results (MAP)

	1 FbackImg	2 FbackImg	3 FbackImg
$noFback$	0.013	0.013	0.013
$eFus$	0.066	0.082	0.085
$lFus$	0.066	0.082	0.085
$rrText$	0.055	0.069	0.075
$rrVis$	0.034	0.036	0.031
$trMed$	0.061	0.078	0.081
$tOnlyFix$	0.063	0.080	0.081
$vOnlyFix$	0.041	0.060	0.070
$tOnlyAd$	0.085	0.095	0.112
$vOnlyAd$	0.036	0.063	0.081
$r_1, r_2 fixed$	**0.084**	**0.101**	**0.116**
$\mathbf{r_1, r_2 adapt}$	0.091	**0.12***	**0.142***

Let us now add another visual feature, a global colour histogram computed in RGB colour space. In the case of an early fusion model, this extra visual feature will be concatenated with the combined (concatenated) vector of local and text features. Late fusion will naturally incorporate colour histogram as additional aggregation factor. Pre-filtering is going to involve an additional, last step, re-ranking by colour histogram. Similarly for transmedia fusion approach, we add an extra step, aggregation of the colour histogram representations' scores corresponding to the top retrieved images.

The results are presented in Table 6. The hybrid CBIR relevance feedback model with fixed parameters values performed best for $r_1 = 0.2$, $r_2 = 0.2$, $r_3 = 0.8$, where r_1, r_2 correspond to global and local visual features respectively, and r_3 corresponds to the textual feature.

TABLE VI: Simulated Relevance Feedback, ImageCLEF2007photo results (MAP) with additional visual feature

	1 FbackImg	2 FbackImg	3 FbackImg
$noFback$	0.013	0.013	0.013
$eFus$	0.066	0.082	0.085
$lFus$	0.066	0.082	0.085
$rrText$	0.053	0.068	0.075
$rrVis$	0.034	0.035	0.032
$trMed$	0.064	0.080	0.083
$r_1, r_2, r_3 fixed$	**0.088**	**0.107**	**0.120**
$\mathbf{r_1, r_2, r_3 adapt}$	0.093	**0.129***	**0.153***

From the results table we can see that the baselines did not benefit much from the addition of global visual feature. However, both the hybrid CBIR relevance feedback model and our enhanced one recorded an improvement in terms of MAP.

It is evident, that the query and its context can be correlated to a different extent. We can utilize the information about this relationship strength to automatically adjust the weights corresponding to query and its context in relevance feedback.

Thus, each query (visual example) can be associated with a particular combination of weights, unique for this query.

V. Conclusions

It has been shown that query can be correlated with its context to a different extent. Inspired by this observation, we incorporate an adaptive weighting scheme into the hybrid CBIR relevance feedback model. Thus, each query is associated with unique set of weights corresponding to the relationship strength between visual query and its visual context as well as the textual query and its textual context. The higher the number of terms or visual terms (mid-level features) co-occurring between current query and the context, the stronger the relationship and vice versa. If the relationship between query and its context is weak, context becomes important. We adjust the probability of the original query terms; the adjustment will significantly modify the original query. If the aforementioned relationship (similarity) between query and its context is strong, however, context will not help much. The original query terms will tend to dominate the whole term distribution in the modified model. The adjustment will not significantly modify the original query.

We tested the enhanced model within the user simulation framework. For fair comparison purposes, the best performing sets of fixed weights were selected. We have shown that our enhanced model with adaptive weighting scheme can outperform the original one with fixed weights. Moreover, an addition of another visual feature (colour histogram, global feature) further improved both the hybrid CBIR relevance feedback model and the enhanced model's performance, whereas the performance of the baselines did not change much.

VI. Future Work

The future work will involve testing different notions of correlation within the proposed framework. In this paper, we incorporate document/image level correlations only. We can also consider a visual counterpart to Hyperspace Analogue to Language, where a window of a fixed size (e.g. square, circular) is shifted from one instance of a visual word to another.

We have recently switched to 1 million image collection and are going to continue increasing the size of our data-set.

Acknowledgment

This work has been partially funded by the EC FP7 OPENi project no. 317883 on Internet of Services (http://www.openi-ict.eu/). The motivation of the project is to provide cloud-based applications and services for mobile users. The AmbieSense work is focused on providing advanced search facilities for large-scale multimedia and social media applications in the cloud.

References

[1] N. Bhowmik, V. R. Gonzalez, V. Gouet-Brunet, H. Pedrini, G. Bloch. Efficient fusion of multidimensional descriptors for image retrieval. *IEEE International Conference on Image Processing*, 5766–5770, 2014.

[2] Y.C. Chang, H.H. Chen. Increasing relevance and diversity in photo retrieval by result fusion. *Working Notes of CLEF 2008*, 2008.

[3] S. Clinchant, J. Ah-Pine, G. Csurka. Semantic combination of textual and visual information in multimedia retrieval. *In ACM International Conference on Multimedia Retrieval (ICMR)*, 2011.

[4] A. Depeursinge, H. Muller. Fusion techniques for combining textual and visual information retrieval. *ImageCLEF, The Springer International Series on Information Retrieval*, 32:95–114, 2010.

[5] A. Goker. Context learning in Okapi. *Journal of Documentation*, 80–83, 1997.

[6] D. He, A. Goker, D.J. Harper. Combining Evidence for Automatic Web Session Identification. *Information Processing and Management Journal, Special Issue on "Context in Information Retrieval"*, 38:5, 605–742, 2002.

[7] A. Goker, H. Myrhaug, R. Bierig. Context and Information Retrieval, in Information Retrieval: Searching in the 21st Century. *John Wiley and Sons*, 2009.

[8] M. Grubinger, P. Clough, A. Hanbury, H. Muller. Overview of the ImageCLEF 2007 photographic retrieval task. *Working Notes of the 2007 CLEF Workshop*, 2007.

[9] L. Kaliciak, D. Song, N. Wiratunga, J. Pan. Novel local features with hybrid sampling technique for image retrieval. *Proceedings of Conference on Information and Knowledge Management (CIKM)*,1557–1560 , 2010.

[10] L. Kaliciak, D. Song, N. Wiratunga, J. Pan. Combining visual and textual systems within the context of user feedback. *19th International Conference on Multimedia Modeling*, 7732(1):445–455, 2013.

[11] L. Kaliciak, H. Myrhaug, A. Goker, D. Song. On the duality of specific early and late fusion strategies. *Information Fusion (FUSION), 17th International Conference on*, 1–8, 2014.

[12] T. Mensink, G. Csurka, F. Perronnin. LEAR and XRCE's participation to visual concept detection task - ImageCLEF 2010. *Proceedings of the 14th Annual ACM International Conference on Multimedia*, 77–80, 2010.

[13] T. Mensink, J. Verbeek, G. Csurkay. Weighted transmedia relevance feedback for image retrieval and auto-annotation. *Technical Report Number 0415*, 2011.

[14] E. Nowak, F. Jurie, B. Triggs. Sampling strategies for bag-of-features image classification. *Lecture Notes in Computer Science*, 3954(490), 2006.

[15] D. Picard, N. Thome, and M. Cord. An efcient system for combining complementary kernels in complex visual categorization tasks. *Proceedings of 17th IEEE International Conference on Image Processing*, 3877-3880, 2010.

[16] V. Risojevic and Z. Babic. Fusion of global and local descriptors for remote sensing image classication. *IEEE Geoscience and Remote Sensing Letters*, 10(4):836–840, 2013.

[17] J. Teevan, S. Dumais, E. Horvitz. Personalizing Search via Automated Analysis of Interests and Activities. *28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 449–456, 2005.

[18] J. Wang, D. Song, L. Kaliciak. Tensor product of correlated text and visual features: a quantum theory inspired image retrieval framework. *AAAI-Fall 2010 Symposium on Quantum Information for Cognitive, Social, and Semantic Processes*, 109–116, 2010.

[19] X. Wang, M. Yang, H. Qi, S. Li, and T. Zhao. Adaptive Weighting Approach to Context-Sensitive Retrieval Model. *8th Asia Information Retrieval Societies Conference*, 7675:417–426, 2012.

[20] S. Wu, Y. Xing, J. Li, and J. Bi. Adaptive Data Fusion Methods for Dynamic Search Environments. *8th Asia Information Retrieval Societies Conference*, 7675:336–345, 2012.

[21] W. Zhang, Z. Qin, and T. Wan. Image scene categorization using Multi-Bag-of-Features. *Proceedings of International Conference on Machine Learning and Cybernetics*, 4:1804–1808, 2011.

[22] ImageCLEF website. *www.imageclef.org*