Content-based retrieval of visual information

Oerlemans, A.A.J.

Chapter 7

Multi-Dimensional Maximum Likelihood

7.1 Introduction

In retrieval applications, there is a point where the similarity between two documents has to be determined. Documents similar to the query should be ranked higher in the search results than documents that are less similar.

An interesting question to ask is: how do we determine similarity? When are two image features similar? Assuming that these features are represented by vectors of real numbers, standard ways of determining feature similarity exist, such as the sum of absolute distances (SAD). However, these methods assume certain properties of the data that is used. For example, as will be described later in this chapter, using a sum of squared distances (SSD) as a similarity metric, assumes that the feature differences of similar images are normally distributed.

Previous research has shown that this is not always true and that more optimal similarity measures exist if the distribution of feature data of similar images is known. This research extends this work to a multi-dimensional situation where not only the distribution of feature value differences is analyzed, but also the relation of these differences to the feature values themselves. In other words, given a set of similar images, we are not examining the distribution of $(x - y)$ for each pair of features but the distribution of (x, y), which is a 2D distribution. This approach results in the multi-dimensional maximum likelihood (MDML) similarity measure.

7.2 Definitions

A distance is a function D with nonnegative real values, defined on the Cartesian product $X \times X$ of a set X. So, for every $x, y \in X$:

1. $D(x, y) = 0$

A distance is called a metric if the following properties also hold for every $x, y, z \in X$:

2. $D(x, y) = 0$ if and only if $x = y$
3. $D(x, y) = D(y, x)$
4. $D(x, z) = D(x, y) + D(y, z)$

A set X provided with a metric is called a metric space. As an example, every set X has the trivial discrete metric $D(x, y) = 0$ if $x = y$ and $D(x, y) = 1$ otherwise.

If one of the metric conditions is not met, the distance function is referred to as a similarity measure.

7.3 Detailed description

We begin by reviewing the theory of maximum likelihood theory. Given two images X and Y with feature vectors x and y, the probability of these two being similar, is the product of the probabilities of the similarity of each element of the vector.

$$Sim(x, y) = \prod_{i=1}^{n} P_{sim}(x_i, y_i) \tag{7.1}$$

These individual probabilities $P_{sim}(x_i, y_i)$ are directly linked to the distribution of the feature value co-occurrences and are often modeled by a chosen probability density function.

The origin of the widely used L_2 distance metric is the assumption that the differences between feature vector elements are normally distributed [78], with the same parameters for each element. This results in the following similarity metric:

$$Sim(x, y) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x_i - y_i)^2}{2\sigma^2}\right)} \tag{7.2}$$

If one wants to find the most similar image to an image X, we loop through all images Y to find the image with the highest similarity value resulting from 7.2.

However, since σ and μ are both constants in this formula, we can simplify the maximization problem using:

$$Sim(x,y) = \prod_{i=1}^{n} e^{-(x_i - y_i)^2} \tag{7.3}$$

Applying $ln(ab) = ln(a) + ln(b)$ to this function yields:

$$Sim(x,y) = \sum_{i=1}^{n} ln\left(e^{-(x_i - y_i)^2}\right) \tag{7.4}$$

Which in turn can be simplified to:

$$Sim(x,y) = \sum_{i=1}^{n} -(x_i - y_i)^2 \tag{7.5}$$

Finally, if we convert the maximization problem to a minimization problem, we can use:

$$Sim(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2 \tag{7.6}$$

This is exactly the sum of squared distances. So now we can conclude that if the differences of all feature values have the same normal distribution, using L_2 as a distance metric is justified.

As mentioned before, the normal distribution does not always represent the true distribution in retrieval experiments. For example in motion detection in video, a representative distribution and the best fit Gaussian is shown in figure 7.1.

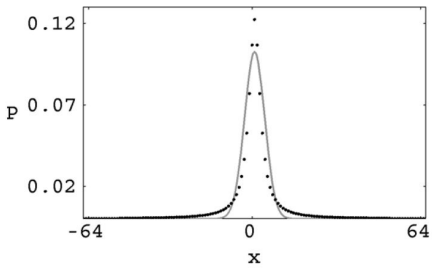

Figure 7.1: An example of the best fit Gaussian to the true distribution in one of our video retrieval experiments.

Especially the more prolonged tails of the true distribution are noticeable. This suggests that other distributions might have a better fit to the true underlying distribution and that other similarity metrics should be used in this case.

There are three assumptions that are made when the L_2 distance metric is used for determining the similarity of two feature vectors:

1. The difference $(x - y)$ between elements of the feature vectors contains all information needed to determine similarity.
2. The differences between the elements of similar feature vectors follow a Gaussian distribution.
3. All differences of the elements in the feature vector have identical independent distributions.

The first assumption is widely used in all types of distance metrics. It is assumed that the differences of feature vectors follow a certain distribution for similar images and that this difference captures it all. However, there are many other possibilities that can be analyzed. Instead of using the differences, looking at the absolute values (x, y) or even $(x + y)$ or $(x * y)$ might result in better distance metrics.

The second assumption states that $(x-y)$ follows a Gaussian distribution, which is explained by the derivation of the L_2 metric in this chapter. The third assumption is based on the fact that the same distance metric is used for each feature value when determining the similarity of feature vectors. The same function is used for each feature vector element.

If the third assumption is not true, then the distribution of each feature vector element of similar images should be determined and a suitable distance should be selected based on that distribution. This will not be addressed in this chapter, but it will be the focus of future research.

If the second assumption is not true (which it is not, as was demonstrated by previous research [76]), a better approximation for the true distribution should be used to select a distance metric that better suits the data. This will be the focus of the rest of this chapter.

If the first assumption is not true, more information should be extracted from the original feature values. In this chapter, we do this by examining the distribution of (x, y).

7.4 Related work

A lot of research has been done in the field of distances and metrics. Below is a review of a few recent papers. Almost all try to improve on the standard L_1 and L_2 norms by looking at their general form, the Minkowski L_p norm:

$$D_{L_p}(x, y) = \left(\sum (x - y)^p\right)^{\frac{1}{p}} \tag{7.7}$$

With SAD (L_1) and SSD (L_2) being two instances of this norm.

Sebe et al. [78] analyze the true noise distributions of similar items from various test sets and the authors conclude that the implied Gaussian or Exponential distributions are often not close enough to the true noise distributions. The presented Cauchy metric has increased performance over both L1 and L2. Further research, also by Sebe et al. [76], shows that using the histogram of feature value differences of similar items, further increases the similarity measure performance. This research addresses assumption 2, the differences of feature values of similar images do not follow a Gaussian distribution.

Yu et al. [97] build on the research by Sebe and introduce several other distance measures based on different probability distributions that are implied by the harmonic mean and the geometric distribution, as opposed to the arithmetic mean and the median that are related to L_2 and L_1. Yu also correctly identifies the problem of the possibility of having different noise distributions for individual feature elements. He shows that using several distance measures for sets of feature elements, by using a boosting algorithm to find these sets, the performance of the distance measure increases. This research addresses assumptions 2 and 3, the differences of feature values of similar images do not follow a Gaussian distribution and also the distributions of differences of feature elements are not the same for each element.

However, the second assumption, stating that $(x - y)$ contains all information needed to determine similarity, is not addressed by these papers. Our research will focus on this assumption to determine if more information should be used.

7.5 Multi-Dimensional Maximum Likelihood similarity (MDML)

As mentioned above, previous research [76] has shown that analyzing the distribution $P(x_i - y_i)$ of feature differences for adjusting the similarity metric results in better retrieval results. The resulting similarity metric was called the maximum likelihood metric. In this research we are directing our focus at the distribution of $P(x_i, y_i)$ and we attempt to create a 2D version of this metric. Using (x, y) seems like the most general approach to analyze the data.

Recall that a similarity measure can be thought of as a histogram of feature value co-occurrences. If given enough training samples, the normalized histogram will give a representation of the true joint probability of the occurrence of each feature value pair for a certain class of images. The similarity of two images, given the class C of one of them, can then be calculated by

$$Sim(x, y) = \prod_{i=1}^{n} H_C(x_i, y_i) \tag{7.8}$$

Where H_C is the probability of the two feature values occurring for similar images, determined by the histogram for class C. To convert this into a minimization problem and to get more numerical stability in the calculations (by converting the product to a sum), we get:

$$Sim(x, y) = -\sum_{i=1}^{n} log\left(H_C(x_i, y_i)\right) \tag{7.9}$$

Determining the similarity of two feature vectors is now reduced to directly using the true 2D distribution of similar feature values.

7.6 Experiments on stereo matching

Stereo matching is the process of finding corresponding image locations in sets of images, such that these locations represent the same real-world location. The difference between the pixel locations in the images is called the disparity.

Stereo matching algorithms come in various forms, but three important categories can be distinguished: local, semi-global and global methods. Local methods try to find the best match to a single location by using information around that location or possibly even only the pixel itself. Global methods try to optimize a global correspondence function, which causes local correspondences to be influenced by neighboring locations. Semi-global methods use the same principle as global methods, a form of functional optimization, however restricted to a subset of the global set of correspondences to improve calculation times.

A well-known dataset for testing stereo matching algorithms is the Middlebury dataset. A total of 28 sets of stereo images and corresponding ground truth were released between 2001 and 2006. In our experiments we have used the datasets released in 2001, 2003 and a few from 2005, together with two other sets (Map and Tsukuba) that were used in the experiments from [70].

7.6.1 Results - template based

For our first experiments, we have selected a template size of 11 and a search range of one line, because the images are already rectified. Table 7.1 shows the accuracy of template based stereo matching for each Middlebury dataset and various similarity measures. The experiments were carried out with 10-fold cross validation.

7.6.2 Results - pyramidal template based

These experiments were based on a variation of the template based stereo matching algorithm: it uses a 5-layer pyramid of subsampled versions of the original

Dataset	L1	L2	ML	MDML
Barn 1	93.88 (0.07)	92.93 (0.05)	94.75 (0.08)	96.66 (0.05)
Barn 2	91.43 (0.09)	91.70 (0.12)	94.46 (0.10)	95.63 (0.09)
Bull	95.27 (0.07)	95.54 (0.08)	97.26 (0.08)	97.66 (0.04)
Poster	84.58 (0.13)	85.29 (0.09)	87.08 (0.11)	94.18 (0.22)
Sawtooth	93.92 (0.09)	92.97 (0.11)	94.90 (0.15)	96.59 (0.13)
Venus	92.46 (0.13)	92.45 (0.14)	93.75 (0.14)	95.53 (0.10)
Map	92.30 (0.09)	91.33 (0.06)	92.57 (0.05)	93.05 (0.09)
Tsukuba	51.27 (1.16)	50.11 (1.18)	51.84 (0.97)	71.05 (0.87)
Cones	81.23 (0.26)	81.77 (0.24)	88.72 (0.21)	92.14 (0.16)
Teddy	77.97 (0.26)	77.70 (0.21)	84.40 (0.22)	88.68 (0.24)
Art	56.01 (0.13)	54.87 (0.16)	68.10 (0.19)	73.01 (0.20)
Books	72.83 (0.36)	75.43 (0.20)	78.01 (0.26)	86.65 (0.26)
Dolls	78.63 (0.12)	76.51 (0.11)	81.69 (0.11)	85.90 (0.16)

Table 7.1: Average accuracy and unbiased standard deviation of found correspondences for the Middlebury datasets and various similarity measures

image. Searching for a match starts at the lowest resolution layer. Each layer has a separately trained ML or MDML classifier. The matching template location on the low resolution image is used as the center of the limited search range for the next higher resolution.

For these experiments, we have selected a template size of 11 and a search range of one line, because the images are already rectified. Table 7.2 shows the accuracy of template based stereo matching for each Middlebury dataset and various similarity measures. The experiments were carried out with 10-fold cross validation.

Dataset	L1	L2	ML	MDML
Barn 1	93.88 (0.07)	92.93 (0.05)	94.75 (0.08)	96.66 (0.05)
Barn 2	90.90 (0.29)	91.20 (0.33)	91.50 (0.22)	95.58 (0.11)
Bull	94.94 (0.06)	95.15 (0.08)	95.13 (0.07)	97.79 (0.04)
Poster	85.48 (0.14)	86.03 (0.09)	85.84 (0.18)	94.41 (0.20)
Sawtooth	93.03 (0.10)	91.98 (0.11)	93.44 (0.13)	96.65 (0.13)
Venus	92.51 (0.10)	92.10 (0.09)	86.61 (0.09)	96.12 (0.01)
Map	90.39 (0.10)	88.69 (0.10)	91.45 (0.08)	94.53 (0.16)
Tsukuba	46.22 (1.01)	41.23 (0.75)	51.64 (0.65)	74.61 (0.09)
Cones	75.91 (0.27)	75.63 (0.22)	80.29 (0.28)	92.14 (0.17)
Teddy	76.24 (0.26)	75.57 (0.16)	83.82 (0.28)	89.64 (0.24)
Art	53.73 (0.23)	51.44 (0.20)	64.53 (0.28)	75.48 (0.10)
Books	68.27 (0.36)	68.49 (0.21)	75.72 (0.15)	87.99 (0.29)
Dolls	74.45 (0.17)	72.00 (0.18)	75.28 (0.14)	87.03 (0.12)

Table 7.2: Average accuracy and unbiased standard deviation of found correspondences for the Middlebury datasets and various similarity measures

7.7 Future work

Given the MDML histogram and assuming enough training samples were available to find the true probability distribution, an optional next step is to find a parametric representation of the distribution. This will reduce the need to keep the histogram data for calculating distances. Several methods can be thought of when converting a histogram to a 2D function:

- PCA
- Surface approximation
- Wavelets

As the research by Sebe has shown, using an approximation to the true distribution will result in suboptimal accuracy, but usually this will be compensated for by processing speed or memory consumption.

In case of the stereo matching algorithm described in this chapter, the 2D histogram uses 64 kilobytes of memory and uses a lookup in the probability distribution array for each pixel in the template matching step. Future research will be focusing on the accuracy of stereo matching with parametric 2D approaches based on the true distribution, taking into account the effect on processing time and memory.