PORTMANTEAU VOCABULARIES FOR MULTI-CUE IMAGE REPRESENTATIONS

FAHAD SHAHBAZ KHAN[1], JOOST VAN DE WEIJER[1], ANDREW D. BAGDANOV[1,2], MARIA VANRELL[1]
[1]CENTRE DE VISIO PER COMPUTADOR, COMPUTER SCIENCE DEPARTMENT
[1]UNIVERSITAT AUTONOMA DE BARCELONA, EDIFÍCI O, CAMPUS UAB (BELLATERRA), BARCELONA, SPAIN
[2] MEDIA INTEGRATION AND COMMUNICATION CENTER, UNIVERSITY OF FLORENCE, ITALY

PROBLEM STATEMENT

Goal: How to construct efficient-multi cue vocabularies for large-scale data sets?
Problems: Existing fusion approaches are problematic for data sets with several hundred object categories.

Method	Cue-Binding	Cue-Weighting	Scalability
Early Fusion	Yes	Hard	Yes
Late Fusion	No	Yes	Yes
Color Attention[2]	Yes	Yes	No

Desired Properties:
Cue-Binding: This property refers to combining color and shape information at the local feature level. This allows for the description of blue corners, red blobs, etc.
Cue-Weighting: This implies constructing a separate visual vocabulary for both color and shape. Having this property allows for efficient cue-weighting.
Scalability: The final dimensionality should be independent of number of object categories in a data set.

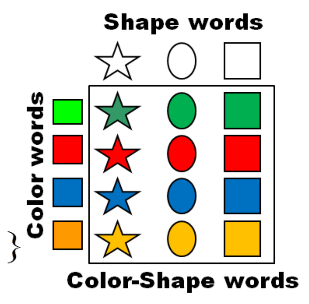
PRODUCT VOCABULARIES

A simple way to ensure cue-binding is by a product vocabulary of primitive visual cues.

Shape vocabulary: $S = \{s_1, s_2, ..., s_M\}$

Color vocabulary: $C = \{c_1, c_2, ..., c_N\}$

Product vocabulary: $W = \{w_1, w_2, ..., w_T\}$
$= \{\{s_i, c_j\} \mid 1 \le i \le M, 1 \le j \le N\}$
$T = M \times N$



Drawbacks: Product vocabularies are very high dimensional. The resulting representation leads to overfitting on the training set.

COMPACT VOCABULARIES USING DITC

To obtain compact representations, the DITC algorithm[1] is used to compress visual vocabularies. The algorithm is designed to find fixed number of clusters. The DITC optimizes a global objective function:

$$I(R,W) - I(R,W^R) = \sum_{j=1}^{J} \sum_{w_t \in W_j} p(w_t) KL\left(p(R \mid w_t), p(R \mid W_j)\right)$$

Cluster words containing similar *mutual information* to classes | word priors | similarity between distributions

DITC iteratively optimizes the above objective function:

1. Compute the cluster distributions according to:
$p(R \mid W_j) = \sum_{w_t \in W_j} p(w_t) p(R \mid w_t)$
2. Re-assign the words to the clusters based on their closeness in KL-divergence respectively:
$j^*(w_t) = \arg\min_j KL(p(R \mid w_t), p(R \mid W_j))$

PORTMANTEAU VOCABULARIES

Compress product vocabularies using the DITC technique. This results in a compact multi-cue visual vocabulary which is used to construct a color-shape histogram.

$$I(R,W) - I(R,W^R) = \sum_{j=1}^{J} \sum_{w_t \in W_j} p(w_t) KL\left(p(R \mid w_t), p(R \mid W_j)\right)$$

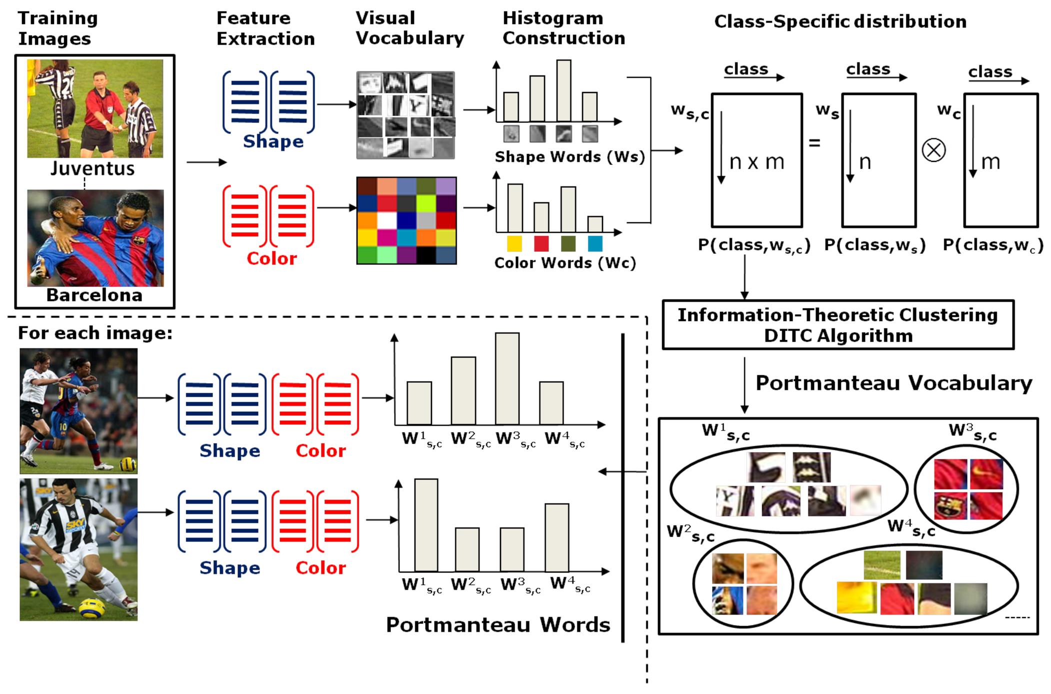
Joint color-shape distributions | Portmanteau words

Novelty: The DITC algorithm is not investigated before to handle the problem of multi-cue visual vocabularies.

OUR APPROACH: PORTMANTEAU VOCABULARIES

Procedure:

1. Construct separate color and shape vocabularies.
2. Empirical class-conditional word distributions of color and shape using the training set.
3. Estimate joint cue distribution assuming conditional independence over classes.
4. Compress the large product vocabularies using the DITC algorithm to obtain Portmanteau words.
5. A new color-shape histogram is constructed by using the new index list output by DITC.

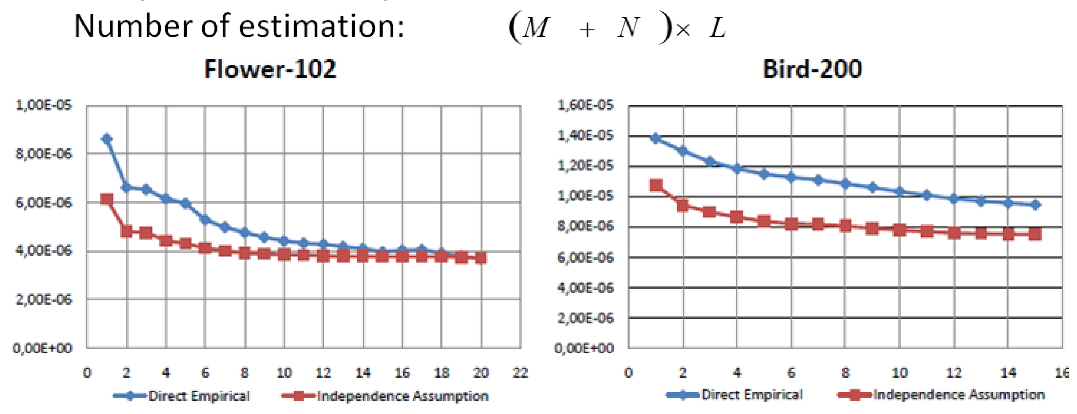


ESTIMATING JOINT-CUE DISTRIBUTIONS

Observation: Modeling joint-cue distributions independently over the class is statistically more robust than empirical dependent joint-distribution directly.

Class conditional probability: $p(w \mid R) = p(s, c \mid R)$
Number of estimation: $M \times N \times L$

Independence assumption: $p(w \mid R) \propto p(s \mid R) p(c \mid R)$
Number of estimation: $(M + N) \times L$



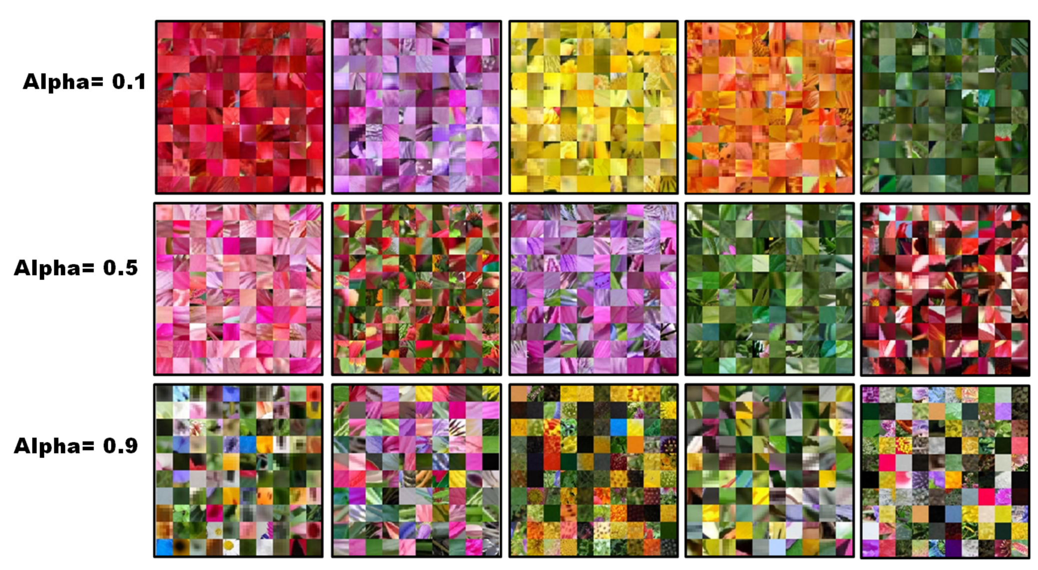
Jenson-Shannon divergence between each estimate and the true joint distribution.

1. Results are provided as a function of number of training images.
2. Low JS means a better estimate of the true joint-cue distribution.
3. Results shows that independence assumption yields similar of better estimates than empirical counterparts.

CUE-WEIGHTING

The independence assumption additionally allows for efficient weighting of cues [0,1]:

$$p^{\alpha}(s,c \mid R) \propto p(s \mid R)^{\alpha} p(c \mid R)^{1-\alpha}$$

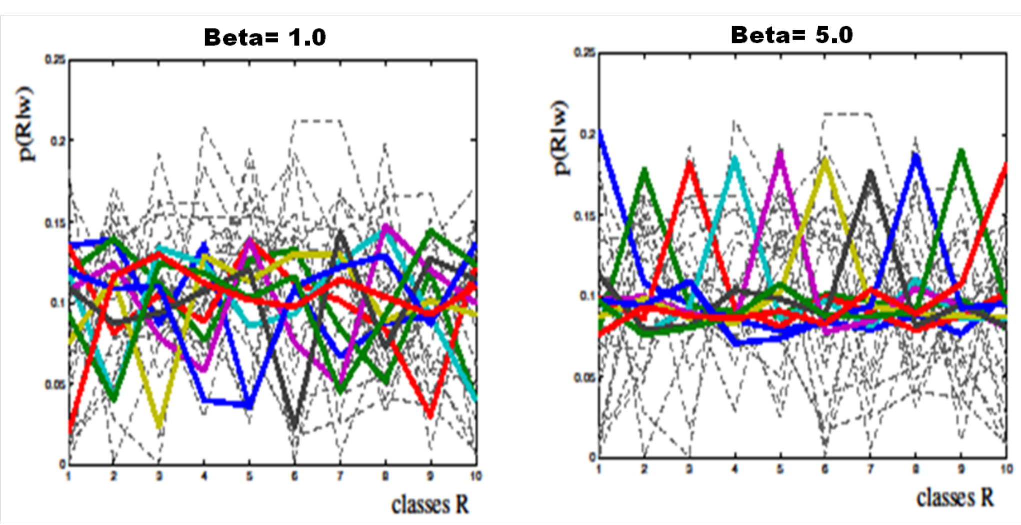


The effect of weighting on Portmanteau clusters.

HIGHLY DISCRIMINATIVE CLUSTERS

The beta parameter directs the DITC to find clusters discriminative for a single category:

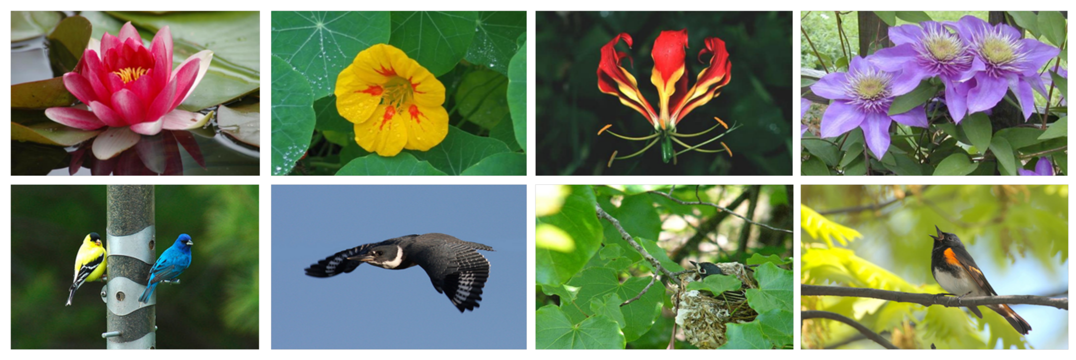
$$p^{\alpha,\beta}(s,c \mid R) \propto \left(p(s \mid R)^{\alpha} p(c \mid R)^{1-\alpha}\right)^{\beta}$$



The effect of beta on DITC clusters. A higher beta directs DITC to construct Portmanteau each discriminating one class.

EXPERIMENTAL VALIDATION

We validate our approach on two difficult data sets Bird-200 (6000 images) and Flower-102 (8000 images).



Method	Flower-102	Bird-200
Shape Only	60.7	12.9
Color Only	48.5	16.8
Early Fusion	70.5	17.0
Direct Empirical	64.6	18.9
Independent	63.5	19.8
Independent + α	66.4	21.6
Independent + α **+** β	**73.3**	**22.4**

Comparison with the state-of-the-art: Our approach yields competitive results by only combining two cues.

Method	Flower-102	Bird-200
OpponentSIFT	69.2	14.0
C-SIFT	65.9	13.9
MKL [Nilsbeck08]	72.8	-
MKL [Branson10]	-	19.0
Random Forest	-	19.2
Saliency	71.0	-
Our Approach	**73.3**	**22.4**

CONCLUSIONS

1. We propose a new method to construct multi-cue vocabularies.
2. We compress product vocabularies to construct discriminative compound visual words.
3. Assuming independence of cues given the class provides robust estimation.
4. Additionally it allows for efficient cue-weighting.
5. Our final representation is compact, maintains cue binding and admits cue weighting.

REFERENCES

[1] Inderjeet S. Dhillon, Subramanyam Mallela and Rahul Kumar. A divisive information-theoretic clustering algorithm for text classification. *JMLR '03*

[2] Fahad S. Khan, Joost van de Weijer and Maria Vanrell. Top-down color attention for object recognition. In *ICCV '09*