

Recognizing image “style” and activities in video using local features and naive Bayes – Reading Notes

1. Main goal of the paper

The paper proposes a simple framework for classifying images and videos by their “style” or “type”, such as identifying an artist from a painting or recognizing an activity in a video sequence. It relies on local DCT-based features extracted from small spatial or spatio-temporal blocks and uses a naive Bayes classifier to decide the class of each block and then of the whole image/video.

2. Key methods and ideas

- Images are divided into overlapping 9×9 blocks (and $5 \times 5 \times 5$ blocks in video), converted to gray-scale and normalized.
- DCT coefficients of each block are treated as binary features, by thresholding their absolute values per artist/class to maximize mutual information.
- Naive Bayes is trained per class pair, selecting a small set of highly informative DCT features, and classification is done by majority vote over blocks, producing both a global label and a pixel-wise style/activity map.
- The same idea is extended to video by applying a 3D DCT on spatio-temporal blocks around each pixel (stationary camera assumption).

3. Experimental results – short summary

- Artist classification: five painters (Rembrandt, Van Gogh, Picasso, Magritte, Dali) with about 10 training images each, reaching around 86% success in a tournament scheme.
- Old vs. new photographs: separation between 19th-century photographs and digital camera images using the same local DCT-Bayes approach.
- Video activity: low-resolution (64×64) sequences of walking vs. hand-waving, using $5 \times 5 \times 5$ DCT blocks, achieving high pixel-wise classification rates for both activities.

4. Related work (2–3 complementary references)

- Texture / style via local statistics: de Bonet & Viola (1998) use non-parametric multi-scale models for texture recognition, emphasizing statistics of filter responses instead of explicit features.
- Event-based / temporal texture analysis: Zelnik-Manor & Irani (2001) analyze events in video via temporal textures and gradient distributions, providing a more global, histogram-based alternative to Keren’s very local blocks.
- Deep learning for activity recognition: modern works (e.g., 3D CNNs and transformer-based models for video) use learned spatio-temporal features and typically outperform hand-crafted DCT+naive Bayes, at the cost of complexity and data requirements.

5. Initial insights from the readings

- The naive Bayes + DCT approach is extremely simple and efficient, and fits well within a 13-week course project.
- The method is very local and can naturally produce pixel-wise style/activity maps, which matches the requirement to highlight motion regions in color.
- Limitations include: sensitivity to block size and thresholds, assumption of a stationary camera, and relatively shallow modeling of temporal dynamics compared to modern deep methods.
- The related literature suggests two natural directions for improvement: (1) multi-scale / multi-resolution analysis, and (2) better temporal modeling for activities.

6. Planned approach for the course project

- Phase 1–3: Reproduce the original pipeline: gray-scale conversion, block extraction, DCT coefficients, mutual-information-based thresholding, naive Bayes training, and majority-vote classification for both images and simple activity videos.

- Phase 4: Extend the method to classify motion types in the instructor's videos (e.g., translation, rotation, zoom) using 3D DCT features and per-pixel spatio-temporal blocks.
- Phase 5 (improvements):
 - Add a multi-scale variant (different block sizes and/or pyramid) and compare performance.
 - Experiment with simple temporal descriptors (e.g., frame differences or optical-flow magnitude) combined with the DCT-Bayes classifier.
 - Apply basic spatial/temporal smoothing (e.g., majority filtering over neighboring labels) to get cleaner motion regions.
- Phase 6–7: Build a full, reproducible pipeline in GitHub that takes the instructor's videos, classifies motion type over time, and produces color-coded motion region visualizations suitable for the final live demo.