Multimodal classification of birds

Seed-grant proposal

Arnav Bhavsar
Dileep A. D.
Padmanabhan Rajan

Multimedia Analytics and Systems Lab
School of Computing and Electrical Engineering

September 21, 2015

Overview

The objective
The acoustics
The image/video
The machine learning
The budget and other details

Figure: Slaty-headed parakeet. Pic by PPJ.

The objective

- Develop algorithms for automatic analysis of avian biodiversity
- Combine information from acoustic and visual data streams
- Sensors: microphones, cameras
- Apply signal processing and machine-learning techniques to collected data
- Tasks: Species identification, species detection

The objective

- Develop algorithms for automatic analysis of avian biodiversity
- Combine information from acoustic and visual data streams
- Sensors: microphones, cameras
- Apply signal processing and machine-learning techniques to collected data
- Tasks: Species identification, species detection

The objective

- Develop algorithms for automatic analysis of avian biodiversity
- Combine information from acoustic and visual data streams
- Sensors: microphones, cameras
- Apply signal processing and machine-learning techniques to collected data
- Tasks: Species identification, species detection

The objective

- Develop algorithms for automatic analysis of avian biodiversity
- Combine information from acoustic and visual data streams
- Sensors: microphones, cameras
- Apply signal processing and machine-learning techniques to collected data
- Tasks: Species identification, species detection

The objective

- Develop algorithms for automatic analysis of avian biodiversity
- Combine information from acoustic and visual data streams
- Sensors: microphones, cameras
- Apply signal processing and machine-learning techniques to collected data
- Tasks: Species identification, species detection

The objective

- Develop algorithms for automatic analysis of avian biodiversity
- Combine information from acoustic and visual data streams
- Sensors: microphones, cameras
- Apply signal processing and machine-learning techniques to collected data
- Tasks: Species identification, species detection

The motivation

- Birds provide crucial ecosystem services: pollination, seed dispersal, insectivory
- Avian diversity: good indicator of ecosystem health in a local area
- Automatic and semi-automatic sensing devices can be utilized
- Large volume of data captured by these devices
- Algorithms to analyse this data would be useful to ecologists
- Our campus location in the lower Himalays: sensitive ecosystem
- Proposed system can be used for long-term ecological monitoring

The motivation

- Birds provide crucial ecosystem services: pollination, seed dispersal, insectivory
- Avian diversity: good indicator of ecosystem health in a local area
- Automatic and semi-automatic sensing devices can be utilized
- Large volume of data captured by these devices
- Algorithms to analyse this data would be useful to ecologists
- Our campus location in the lower Himalays: sensitive ecosystem
- Proposed system can be used for long-term ecological monitoring

The motivation

- Birds provide crucial ecosystem services: pollination, seed dispersal, insectivory
- Avian diversity: good indicator of ecosystem health in a local area
- Automatic and semi-automatic sensing devices can be utilized
- Large volume of data captured by these devices
- Algorithms to analyse this data would be useful to ecologists
- Our campus location in the lower Himalays: sensitive ecosystem
- Proposed system can be used for long-term ecological monitoring

The motivation

- Birds provide crucial ecosystem services: pollination, seed dispersal, insectivory
- Avian diversity: good indicator of ecosystem health in a local area
- Automatic and semi-automatic sensing devices can be utilized
- Large volume of data captured by these devices
- Algorithms to analyse this data would be useful to ecologists
- Our campus location in the lower Himalays: sensitive ecosystem
- Proposed system can be used for long-term ecological monitoring

The motivation

- Birds provide crucial ecosystem services: pollination, seed dispersal, insectivory
- Avian diversity: good indicator of ecosystem health in a local area
- Automatic and semi-automatic sensing devices can be utilized
- Large volume of data captured by these devices
- Algorithms to analyse this data would be useful to ecologists
- Our campus location in the lower Himalays: sensitive ecosystem
- Proposed system can be used for long-term ecological monitoring

The motivation

- Birds provide crucial ecosystem services: pollination, seed dispersal, insectivory
- Avian diversity: good indicator of ecosystem health in a local area
- Automatic and semi-automatic sensing devices can be utilized
- Large volume of data captured by these devices
- Algorithms to analyse this data would be useful to ecologists
- Our campus location in the lower Himalays: sensitive ecosystem
- Proposed system can be used for long-term ecological monitoring

The motivation

- Birds provide crucial ecosystem services: pollination, seed dispersal, insectivory
- Avian diversity: good indicator of ecosystem health in a local area
- Automatic and semi-automatic sensing devices can be utilized
- Large volume of data captured by these devices
- Algorithms to analyse this data would be useful to ecologists
- Our campus location in the lower Himalays: sensitive ecosystem
- Proposed system can be used for long-term ecological monitoring

The motivation

- Birds provide crucial ecosystem services: pollination, seed dispersal, insectivory
- Avian diversity: good indicator of ecosystem health in a local area
- Automatic and semi-automatic sensing devices can be utilized
- Large volume of data captured by these devices
- Algorithms to analyse this data would be useful to ecologists
- Our campus location in the lower Himalays: sensitive ecosystem
- Proposed system can be used for long-term ecological monitoring

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the individual acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The challenges

- Challenges at various levels
- Acoustics:
  - Complex acoustic environment where recordings are made
  - Overlapping vocalizations, intra-species call variability
  - Background sounds (other animals, human-made sounds, river etc.)
- Image/video:
  - Complex visual environment, visual background clutter
  - Overlapping inter-class visual appearances, local variations
  - Intra-class variations: robustness to changes in pose, motion, light conditions.
- Machine-learning:
  - Fixed-length representations and varying-length representations
  - Dynamic kernels for bird data from different modalities
  - Fusion of modalities
    - Combining the representations from acoustic and image/video modes
    - Combining the decisions from the classifiers for the representations from acoustic and image/video modes
  - Bird indexing and retrieval

The acoustics (cont'd)

- Processing of human speech: techniques can be adapted for birdcalls
- Production mechanisms are different, but have similarities (eg. formant structure)
- Existing techniques include:
  - spectral representations [1],
  - Mel frequency cepstra [2],
  - hidden Markov models [3],
  - sparse representations [4]

[1] H. Tyagi et. al., "Automatic identification of bird calls using spectral ensemble average voice prints", Proc. EUSIPCO, 2006

[2] M. Graciarena et. al., "Acoustic front-end optimization for bird species recognition", Proc. ICASSP, 2010

[3] M. Graciarena et.al., "Bird species recognition combining acoustic and sequence modeling", Proc. ICASSP, 2011

[4] L. N. Tan et. al. "Evaluation of a sparse representation-based classifier for bird phrase classification", Proc. Interspeech, 2012

The acoustics (cont'd)

- Processing of human speech: techniques can be adapted for birdcalls
- Production mechanisms are different, but have similarities (eg. formant structure)
- Existing techniques include:
  - spectral representations [1],
  - Mel frequency cepstra [2],
  - hidden Markov models [3],
  - sparse representations [4]

[1] H. Tyagi et. al., "Automatic identification of bird calls using spectral ensemble average voice prints", Proc. EUSIPCO, 2006

[2] M. Graciarena et. al., "Acoustic front-end optimization for bird species recognition", Proc. ICASSP, 2010

[3] M. Graciarena et.al., "Bird species recognition combining acoustic and sequence modeling", Proc. ICASSP, 2011

[4] L. N. Tan et. al. "Evaluation of a sparse representation-based classifier for bird phrase classification", Proc. Interspeech, 2012

The acoustics (cont'd)

- Processing of human speech: techniques can be adapted for birdcalls
- Production mechanisms are different, but have similarities (eg. formant structure)
- Existing techniques include:
  - spectral representations [1],
  - Mel frequency cepstra [2],
  - hidden Markov models [3],
  - sparse representations [4]

[1] H. Tyagi et. al., "Automatic identification of bird calls using spectral ensemble average voice prints", Proc. EUSIPCO, 2006

[2] M. Graciarena et. al., "Acoustic front-end optimization for bird species recognition", Proc. ICASSP, 2010

[3] M. Graciarena et.al., "Bird species recognition combining acoustic and sequence modeling", Proc. ICASSP, 2011

[4] L. N. Tan et. al. "Evaluation of a sparse representation-based classifier for bird phrase classification", Proc. Interspeech, 2012

The acoustics (cont'd)

- Processing of human speech: techniques can be adapted for birdcalls
- Production mechanisms are different, but have similarities (eg. formant structure)
- Existing techniques include:
  - spectral representations [1],
  - Mel frequency cepstra [2],
  - hidden Markov models [3],
  - sparse representations [4]

[1] H. Tyagi et. al., "Automatic identification of bird calls using spectral ensemble average voice prints", Proc. EUSIPCO, 2006

[2] M. Graciarena et. al., "Acoustic front-end optimization for bird species recognition", Proc. ICASSP, 2010

[3] M. Graciarena et.al., "Bird species recognition combining acoustic and sequence modeling", Proc. ICASSP, 2011

[4] L. N. Tan et. al. "Evaluation of a sparse representation-based classifier for bird phrase classification", Proc. Interspeech, 2012

The acoustics (cont'd)

- Research focus: subspace representations
- A recording can be represented as a fixed-length vector $\mathbf{x}$
- Can be used for various applications, for eg. removing background sounds before classification:
  - Project $\mathbf{x}$ into a subspace of background sounds, and remove this component from $\mathbf{x}$
- Fixed-length representations: utilized in kernel functions for support vector machines (SVMs)

The acoustics (cont'd)

- Research focus: subspace representations
- A recording can be represented as a fixed-length vector $\mathbf{x}$
- Can be used for various applications, for eg. removing background sounds before classification:
  - Project $\mathbf{x}$ into a subspace of background sounds, and remove this component from $\mathbf{x}$
- Fixed-length representations: utilized in kernel functions for support vector machines (SVMs)

The acoustics (cont'd)

- Research focus: subspace representations
- A recording can be represented as a fixed-length vector $\mathbf{x}$
- Can be used for various applications, for eg. removing background sounds before classification:
  - Project $\mathbf{x}$ into a subspace of background sounds, and remove this component from $\mathbf{x}$
- Fixed-length representations: utilized in kernel functions for support vector machines (SVMs)

The acoustics (cont'd)

- Research focus: subspace representations
- A recording can be represented as a fixed-length vector $\mathbf{x}$
- Can be used for various applications, for eg. removing background sounds before classification:
  - Project $\mathbf{x}$ into a subspace of background sounds, and remove this component from $\mathbf{x}$
- Fixed-length representations: utilized in kernel functions for support vector machines (SVMs)

The acoustics (cont'd)

- Research focus: subspace representations
- A recording can be represented as a fixed-length vector $\mathbf{x}$
- Can be used for various applications, for eg. removing background sounds before classification:
  - Project $\mathbf{x}$ into a subspace of background sounds, and remove this component from $\mathbf{x}$
- Fixed-length representations: utilized in kernel functions for support vector machines (SVMs)

The acoustics (cont'd)

- Research focus: subspace representations
- A recording can be represented as a fixed-length vector $\mathbf{x}$
- Can be used for various applications, for eg. removing background sounds before classification:
  - Project $\mathbf{x}$ into a subspace of background sounds, and remove this component from $\mathbf{x}$
- Fixed-length representations: utilized in kernel functions for support vector machines (SVMs)

The visual: research areas

- Fine-grained classification (of birds): Relatively recent research area ($\geq$ 2010)
- Detection, segmentation and tracking of birds (relatively unexplored): adapting general object detection, segmentation and tracking methods
- Learning visual guidance: inverse problem to classification $\Rightarrow$ Given the classes, find the discriminative features
- Sound source localization (active area in general domains): challenges for birds: less visual motion, background sounds

The visual: research areas

- Fine-grained classification (of birds): Relatively recent research area ($\geq$ 2010)
- Detection, segmentation and tracking of birds (relatively unexplored): adapting general object detection, segmentation and tracking methods
- Learning visual guidance: inverse problem to classification
  $\Rightarrow$ Given the classes, find the discriminative features
- Sound source localization (active area in general domains): challenges for birds: less visual motion, background sounds

The visual: research areas

- Fine-grained classification (of birds): Relatively recent research area ($\geq$ 2010)
- Detection, segmentation and tracking of birds (relatively unexplored): adapting general object detection, segmentation and tracking methods
- Learning visual guidance: inverse problem to classification
  $\Rightarrow$ Given the classes, find the discriminative features
- Sound source localization (active area in general domains): challenges for birds: less visual motion, background sounds

The visual: research areas

- Fine-grained classification (of birds): Relatively recent research area ($\geq$ 2010)
- Detection, segmentation and tracking of birds (relatively unexplored): adapting general object detection, segmentation and tracking methods
- Learning visual guidance: inverse problem to classification
  $\Rightarrow$ Given the classes, find the discriminative features
- Sound source localization (active area in general domains): challenges for birds: less visual motion, background sounds

The visual: research areas

- Fine-grained classification (of birds): Relatively recent research area ($\geq$ 2010)
- Detection, segmentation and tracking of birds (relatively unexplored): adapting general object detection, segmentation and tracking methods
- Learning visual guidance: inverse problem to classification
  $\Rightarrow$ Given the classes, find the discriminative features
- Sound source localization (active area in general domains): challenges for birds: less visual motion, background sounds

The visual: existing work

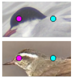
- Fine-grained classification
  - P. Welinder et. al., “Caltech-UCSD Birds 200”, CNS-TR-2010-001. 2010.
  - T. Berg and P. Belhumeur, “POOF: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation”, CVPR 2013.
  - B. Yao et. al., “A codebook-free and annotation-free approach for fine-grained image categorization”, CVPR 2012.
  - L. Xie et. al., “Hierarchical part matching for fine-grained visual categorization”, ICCV 2013.
- Visual guidance: T. Berg and P. Belhumeur, “How do you tell a blackbird from a crow?”, ICCV 2013.
- Detection: D. Song and Y. Xu, “A monocular vision-based low false negative filter for assisting the search for rare bird species using a probable observation data set-based EKF method”, IEEE Trans. Image Processing, 2010.

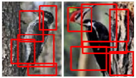
The visual: existing work

- Fine-grained classification
  - P. Welinder et. al., "Caltech-UCSD Birds 200", CNS-TR-2010-001. 2010.
  - T. Berg and P. Belhumeur, "POOF: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation", CVPR 2013.
  - B. Yao et. al., "A codebook-free and annotation-free approach for fine-grained image categorization", CVPR 2012.
  - L. Xie et. al., "Hierarchical part matching for fine-grained visual categorization", ICCV 2013.
- Visual guidance: T. Berg and P. Belhumeur, "How do you tell a blackbird from a crow?", ICCV 2013.
- Detection: D. Song and Y. Xu, "A monocular vision-based low false negative filter for assisting the search for rare bird species using a probable observation data set-based EKF method", IEEE Trans. Image Processing, 2010.

The visual: existing work

- Fine-grained classification
  - P. Welinder et. al., "Caltech-UCSD Birds 200", CNS-TR-2010-001. 2010.
  - T. Berg and P. Belhumeur, "POOF: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation", CVPR 2013.
  - B. Yao et. al., "A codebook-free and annotation-free approach for fine-grained image categorization", CVPR 2012.
  - L. Xie et. al., "Hierarchical part matching for fine-grained visual categorization", ICCV 2013.

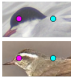


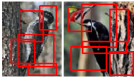


- Visual guidance: T. Berg and P. Belhumeur, "How do you tell a blackbird from a crow?", ICCV 2013.
- Detection: D. Song and Y. Xu, "A monocular vision-based low false negative filter for assisting the search for rare bird species using a probable observation data set-based EKF method", IEEE Trans. Image Processing, 2010.

The visual: existing work

- Fine-grained classification
  - P. Welinder et. al., “Caltech-UCSD Birds 200”, CNS-TR-2010-001. 2010.
  - T. Berg and P. Belhumeur, “POOF: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation”, CVPR 2013.
  - B. Yao et. al., “A codebook-free and annotation-free approach for fine-grained image categorization”, CVPR 2012.
  - L. Xie et. al., “Hierarchical part matching for fine-grained visual categorization”, ICCV 2013.

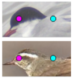
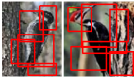




- Visual guidance: T. Berg and P. Belhumeur, “How do you tell a blackbird from a crow?”, ICCV 2013.
- Detection: D. Song and Y. Xu, “A monocular vision-based low false negative filter for assisting the search for rare bird species using a probable observation data set-based EKF method”, IEEE Trans. Image Processing, 2010.

The visual: existing work

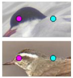
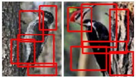
- Fine-grained classification
  - P. Welinder et. al., “Caltech-UCSD Birds 200”, CNS-TR-2010-001. 2010.
  - T. Berg and P. Belhumeur, “POOF: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation”, CVPR 2013.
  - B. Yao et. al., “A codebook-free and annotation-free approach for fine-grained image categorization”, CVPR 2012.
  - L. Xie et. al., “Hierarchical part matching for fine-grained visual categorization”, ICCV 2013.

- Visual guidance: T. Berg and P. Belhumeur, “How do you tell a blackbird from a crow?”, ICCV 2013.
- Detection: D. Song and Y. Xu, “A monocular vision-based low false negative filter for assisting the search for rare bird species using a probable observation data set-based EKF method”, IEEE Trans. Image Processing, 2010.

The visual: existing work

- Fine-grained classification
  - P. Welinder et. al., “Caltech-UCSD Birds 200”, CNS-TR-2010-001. 2010.
  - T. Berg and P. Belhumeur, “POOF: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation”, CVPR 2013.
  - B. Yao et. al., “A codebook-free and annotation-free approach for fine-grained image categorization”, CVPR 2012.
  - L. Xie et. al., “Hierarchical part matching for fine-grained visual categorization”, ICCV 2013.

- Visual guidance: T. Berg and P. Belhumeur, “How do you tell a blackbird from a crow?”, ICCV 2013.
- Detection: D. Song and Y. Xu, “A monocular vision-based low false negative filter for assisting the search for rare bird species using a probable observation data set-based EKF method”, IEEE Trans. Image Processing, 2010.


The visual: possible directions

- Features and frameworks:
  - Patch-based features
  - Body-part features: appearance and geometric relationships
  - Feature learning: deep neural networks, discriminative features
- Frameworks:
  - Sparse representation
  - Markov random fields
  - Hierarchical classification
- Systems:
  - Dataset collection
  - Audio-video systems for monitoring
  - On-board algorithms: detection, tracking

The visual: possible directions

- Features and frameworks:
  - Patch-based features
  - Body-part features: appearance and geometric relationships
  - Feature learning: deep neural networks, discriminative features
- Frameworks:
  - Sparse representation
  - Markov random fields
  - Hierarchical classification
- Systems:
  - Dataset collection
  - Audio-video systems for monitoring
  - On-board algorithms: detection, tracking

The visual: possible directions

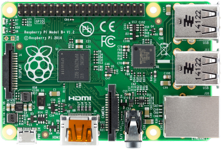
- Features and frameworks:
  - Patch-based features
  - Body-part features: appearance and geometric relationships
  - Feature learning: deep neural networks, discriminative features
- Frameworks:
  - Sparse representation
  - Markov random fields
  - Hierarchical classification
- Systems:
  - Dataset collection
  - Audio-video systems for monitoring
  - On-board algorithms: detection, tracking

The visual: possible directions

- Features and frameworks:
  - Patch-based features
  - Body-part features: appearance and geometric relationships
  - Feature learning: deep neural networks, discriminative features
- Frameworks:
  - Sparse representation
  - Markov random fields
  - Hierarchical classification
- Systems:
  - Dataset collection
  - Audio-video systems for monitoring
  - On-board algorithms: detection, tracking

Example on-field system components

- Data acquisition (audio): rugged, field-deployable recorders
  e.g. Song Meter SM3 recorder from Wildlife Acoustics Inc, USA.
- Data acquisition (video): Network cameras
  e.g. Panasonic WV-SP302
- Processing: Raspberry Pi, Beagle Bone



The machine learning: tasks

- Bird call identification using fixed-length and varying-length acoustic features
- Bird classification from images and videos
- Bird call indexing and retrieval
- Bird image and video indexing and retrieval
- Combining different modalities for classification, indexing and retrieval tasks

The machine learning: tasks

- Bird call identification using fixed-length and varying-length acoustic features
- Bird classification from images and videos
- Bird call indexing and retrieval
- Bird image and video indexing and retrieval
- Combining different modalities for classification, indexing and retrieval tasks

The machine learning: tasks

- Bird call identification using fixed-length and varying-length acoustic features
- Bird classification from images and videos
- Bird call indexing and retrieval
- Bird image and video indexing and retrieval
- Combining different modalities for classification, indexing and retrieval tasks

The machine learning: tasks

- Bird call identification using fixed-length and varying-length acoustic features
- Bird classification from images and videos
- Bird call indexing and retrieval
- Bird image and video indexing and retrieval
- Combining different modalities for classification, indexing and retrieval tasks

The machine learning: tasks

- Bird call identification using fixed-length and varying-length acoustic features
- Bird classification from images and videos
- Bird call indexing and retrieval
- Bird image and video indexing and retrieval
- Combining different modalities for classification, indexing and retrieval tasks

The machine learning: tasks

- Bird call identification using fixed-length and varying-length acoustic features
- Bird classification from images and videos
- Bird call indexing and retrieval
- Bird image and video indexing and retrieval
- Combining different modalities for classification, indexing and retrieval tasks

The machine learning: bird classification

- Classification of birds using SVMs from bird calls and bird images & videos
- The representations for bird call are either fixed-length representation or varying-length representation
- Varying-length representation are either sets of local feature vectors or sequences of local feature vectors
- Dynamic kernel based SVMs for varying-length representation

The machine learning: bird classification

- Classification of birds using SVMs from bird calls and bird images & videos
- The representations for bird call are either fixed-length representation or varying-length representation
- Varying-length representation are either sets of local feature vectors or sequences of local feature vectors
- Dynamic kernel based SVMs for varying-length representation

The machine learning: bird classification

- Classification of birds using SVMs from bird calls and bird images & videos
- The representations for bird call are either fixed-length representation or varying-length representation
- Varying-length representation are either sets of local feature vectors or sequences of local feature vectors
- Dynamic kernel based SVMs for varying-length representation

The machine learning: bird classification

- Classification of birds using SVMs from bird calls and bird images & videos
- The representations for bird call are either fixed-length representation or varying-length representation
- Varying-length representation are either sets of local feature vectors or sequences of local feature vectors
- Dynamic kernel based SVMs for varying-length representation

The machine learning: bird classification

- Classification of birds using SVMs from bird calls and bird images & videos
- The representations for bird call are either fixed-length representation or varying-length representation
- Varying-length representation are either sets of local feature vectors or sequences of local feature vectors
- Dynamic kernel based SVMs for varying-length representation

The machine learning: bird classification (cont'd)

- Some of the dynamic kernels are:
  - GMM-based intermediate matching kernel [5],
  - HMM-based intermediate matching kernel [6],
  - Histogram intersection kernel [7],
  - Spacial pyramid match kernel [8]

[5] A. D. Dileep et. al., "GMM-Based intermediate matching kernel for classification of varying length patterns of long duration speech using SVMs," in IEEE TNNLS, Aug. 2014

[6] A. D. Dileep et. al., "HMM-based intermediate matching kernel for classification of sequential patterns of speech using SVMs," in IEEE TASLP, Dec. 2013

[7] J. C. van Gemert et. al., "Visual word ambiguity," IEEE TPAMI, July 2010

[8] S. Lazebnik et. al., "Beyond bags of features:Spatial pyramid matching for recognizing natural scene categories," in Proceedings of CVPR 2006, June 2006

The machine learning: bird classification (cont'd)

- Some of the dynamic kernels are:
  - GMM-based intermediate matching kernel [5],
  - HMM-based intermediate matching kernel [6],
  - Histogram intersection kernel [7],
  - Spacial pyramid match kernel [8]

[5] A. D. Dileep et. al., "GMM-Based intermediate matching kernel for classification of varying length patterns of long duration speech using SVMs," in IEEE TNNLS, Aug. 2014

[6] A. D. Dileep et. al., "HMM-based intermediate matching kernel for classification of sequential patterns of speech using SVMs," in IEEE TASLP, Dec. 2013

[7] J. C. van Gemert et. al., "Visual word ambiguity," IEEE TPAMI, July 2010

[8] S. Lazebnik et. al., "Beyond bags of features:Spatial pyramid matching for recognizing natural scene categories," in Proceedings of CVPR 2006, June 2006

The machine learning: bird indexing and retrieval

- Matching and retrieval of birds using bird calls and bird images & videos
  - Query-by-example (QBE) based retrieval[9]
  - Query-by-semantics (QBS) based retrieval[10]
  - Query-by-semantic example (QBSE) based retrieval[11]
- Matching and retrieval of birds using kernel methods[12]

[9] A. Marakakis et. al., "Probabilistic relevance feedback approach for content-based image retrieval based on Gaussian mixture models," in IET Image Processing, Feb. 2009

[10] G. Carneiro et. al., "Supervised learning of semantic classes for image annotation and retrieval," in IEEE TPAMI, March 2007

[11] N. Rasiwasia et. al., "Bridging the gap: Query by semantic example," in IEEE Transactions on Multimedia, Aug. 2009

[12] T. Veena, "Image classification, matching and annotation using kernel methods for content based image retrieval for scene images," Ph.D. Thesis, Dept. of CSE, IIT Madras, June 2014.

The machine learning: bird indexing and retrieval

- Matching and retrieval of birds using bird calls and bird images & videos
  - Query-by-example (QBE) based retrieval[9]
  - Query-by-semantics (QBS) based retrieval[10]
  - Query-by-semantic example (QBSE) based retrieval[11]
- Matching and retrieval of birds using kernel methods[12]

---

[9] A. Marakakis et. al., "Probabilistic relevance feedback approach for content-based image retrieval based on Gaussian mixture models," in IET Image Processing, Feb. 2009

[10] G. Carneiro et. al., "Supervised learning of semantic classes for image annotation and retrieval," in IEEE TPAMI, March 2007

[11] N. Rasiwasia et. al., "Bridging the gap: Query by semantic example," in IEEE Transactions on Multimedia, Aug. 2009

[12] T. Veena, "Image classification, matching and annotation using kernel methods for content based image retrieval for scene images," Ph.D. Thesis, Dept. of CSE, IIT Madras, June 2014.

The machine learning: bird indexing and retrieval

- Matching and retrieval of birds using bird calls and bird images & videos
  - Query-by-example (QBE) based retrieval[9]
  - Query-by-semantics (QBS) based retrieval[10]
  - Query-by-semantic example (QBSE) based retrieval[11]
- Matching and retrieval of birds using kernel methods[12]

---

[9] A. Marakakis et. al., "Probabilistic relevance feedback approach for content-based image retrieval based on Gaussian mixture models," in IET Image Processing, Feb. 2009

[10] G. Carneiro et. al., "Supervised learning of semantic classes for image annotation and retrieval," in IEEE TPAMI, March 2007

[11] N. Rasiwasia et. al., "Bridging the gap: Query by semantic example," in IEEE Transactions on Multimedia, Aug. 2009

[12] T. Veena, "Image classification, matching and annotation using kernel methods for content based image retrieval for scene images," Ph.D. Thesis, Dept. of CSE, IIT Madras, June 2014.

The machine learning: multimodal classification and retrieval

- Classfication and retrieval of birds by combining the cues from bird calls and bird images & videos
  - Early fusion: Combining the acoustic, image and video features
  - Late fusion: Combining the decisions from the different classifiers built for bird calls, bird images and bird videos
- Feature selection and combining using multiple kernel learning

The machine learning: multimodal classification and retrieval

- Classfication and retrieval of birds by combining the cues from bird calls and bird images & videos
  - Early fusion: Combining the acoustic, image and video features
  - Late fusion: Combining the decisions from the different classifiers built for bird calls, bird images and bird videos
- Feature selection and combining using multiple kernel learning

The machine learning: multimodal classification and retrieval

- Classfication and retrieval of birds by combining the cues from bird calls and bird images & videos
  - Early fusion: Combining the acoustic, image and video features
  - Late fusion: Combining the decisions from the different classifiers built for bird calls, bird images and bird videos
- Feature selection and combining using multiple kernel learning

Budget and other details

Table: Projected expenses in lakhs INR.

Items	Year 1	Year 2	Year 3	Total
High-end computers (2)	3.0	3.0	0	6.0
Imaging and audio equipment	5.0	3.0	0	8.0
Desktop computers (6)	5.0	0	0	5.0
Contingency	0.5	0.5	1.0	2.0
Travel	0.5	0.5	1.0	2.0
Overall	15.0	8.0	2.0	**23.0**

Future plans: further funding

- **Proposal to SERB:**
  - *Automatic analysis of avian acoustics*.
  - In collaboration with IIT Madras, NCBS and CDAC.
  - Value: Rs 50 lakhs.
  - **Ready for submission.**
- **Proposal planned:** Camera and acoustic sensor networks for a local area (IIT Mandi campus)

Thank you for your attention.

Equipment budget

Table: Equipment budget in thousands INR.

Item	Unit cost	Qty.	Total
Bioacoustic recorder	50	6	300
Network camera	30	5	150
Recorder accessories	15	8	120
DSLR camera and lens	100	1	100
Consumables	50	1	50
Processing hardware	8	5	40
Network access points	3	5	15
Total			**775**