Basketball Training Dataset Collection Strategy

Date: December 13, 2025
Goal: Collect 1,500+ basketball training images from 4 APIs

📸 API Sources & Credentials

1. Pixabay API ✅

- **API Key:** `40138022-57af4b7daf7ed0a81d2f7bded`
- **Source:** Extracted from uploaded screenshot
- **Rate Limit:** 100 requests / 60 seconds (0.6s delay)
- **Max Results:** 200 per page, 500 per query
- **Target:** 400 images
- **Status:** 🔄 In Progress (84.2% complete - 337/400)

2. Pexels API ✅

- **API Key:** `1YOjyRmXxeA5s4q1d7CuxepcnFEgMEhmopApbn3MTS7zPf0vUJsrsQSu`
- **Rate Limit:** 200 requests / hour (18s delay)
- **Max Results:** 80 per page
- **Target:** 400 images
- **Status:** ⏳ Pending

3. Unsplash API ✅

- **Access Key:** `OMJeP8It444nadmsbi5VpbFDfBo9RP0FBwHTLjhulsg`
- **Secret Key:** `gGzB9BbRQfOWfNljeWyj3nBD7X_rEexlqf2sdvFjk_E`
- **Rate Limit:** 50 requests / hour (72s delay)
- **Max Results:** 30 per page
- **Target:** 400 images
- **Status:** ⏳ Pending

4. Kaggle API ✅

- **API Key:** `KGAT_51f015e15f1b1b6313b5e195fe1dd321`
- **Target:** 300 images from basketball datasets
- **Status:** ⏳ Pending

🎯 Search Queries by API

Pixabay Queries (6 queries)

1. `basketball shooting back view`
2. `basketball from behind`

3. `amateur basketball`
4. `youth basketball shooting`
5. `street basketball`
6. `recreational basketball`

Per Query Target: ~66 images (400 ÷ 6)

Pexels Queries (6 queries)

1. `basketball back angle`
2. `basketball player behind`
3. `pickup basketball`
4. `beginner basketball shooting`
5. `basketball training session`
6. `basketball practice`

Per Query Target: ~66 images (400 ÷ 6)

Unsplash Queries (6 queries)

1. `basketball rear view`
2. `basketball training`
3. `college basketball`
4. `casual basketball`
5. `basketball workout`
6. `basketball form`

Per Query Target: ~66 images (400 ÷ 6)

Kaggle Dataset Keywords (3 keywords)

1. `basketball dataset`
2. `basketball training`
3. `basketball players`

Target: ~300 images from top 5 datasets per keyword

📊 Collection Targets

Overall Target: 1,500 images

API	Target	Current	Status	Progress
Pixabay	400	337	🔄 In Progress	84.2%
Pexels	400	0	⏳ Pending	0.0%
Unsplash	400	0	⏳ Pending	0.0%
Kaggle	300	0	⏳ Pending	0.0%
TOTAL	**1,500**	**337**	🔄 **In Progress**	**22.5%**

🗂️ Categorization Strategy

Target Distribution

Category	Current	Target	Need	Priority
Back View	187	250	63	HIGH ⚠️
Good Form	28	228	200	CRITICAL 🔴
Needs Work	15	500	485	CRITICAL 🔴
Poor Form	9	500	490	CRITICAL 🔴
TOTAL	**239**	**1,478**	**1,238**	-

Automatic Categorization Keywords

Back View

- “back view”, “from behind”, “rear view”
- “back angle”, “behind player”
- “back perspective”

Good Form

- “professional”, “perfect form”, “excellent”
- “elite”, “proper technique”
- “flawless”, “textbook”

Needs Work

- “training”, “learning”, “practice”
- “beginner”, “intermediate”
- “developing”, “improving”

Poor Form

- “amateur”, “casual”, “recreational”
- “street”, “pickup”, “youth”
- “inexperienced”

🔧 Technical Implementation

Collection Scripts

1. Main Collection Script

File: `collect_basketball_images.py`
- **Features:**
- Sequential API calls with proper rate limiting
- High-resolution filtering (1080p+)
- Metadata tracking (source, query, dimensions, likes, views)
- Error handling and retry logic
- Progress reporting via tqdm
- **Status:** 🔄 Running in background (PID: 24829)

2. Fast Parallel Collector

File: `fast_parallel_collector.py`
- **Features:**
- ThreadPoolExecutor with 10 concurrent downloads
- Batch processing for faster collection
- Reduced rate limiting for APIs that allow it
- Same metadata tracking as main script
- **Status:** 📦 Ready for use (not currently running)
- **Use Case:** Speed up Pexels/Unsplash collection if needed

3. Progress Monitor

File: `monitor_collection.py`
- **Features:**
- Real-time progress tracking
- Visual progress bars
- Process status checking
- Continuous watch mode (`--watch [seconds]`)
- **Usage:** `python3 monitor_collection.py` or `python3 monitor_collection.py --watch 30`

4. Image Categorizer

File: `categorize_images.py`
- **Features:**
- Automatic keyword-based categorization
- Weighted random distribution for uncategorized images
- Target achievement tracking
- Categorization report generation
- **Status:** 📦 Ready to run after collection completes

📁 Directory Structure

```
/home/ubuntu/basketball_app/training_data/
├── raw_images/                    # Downloaded images by API
│   ├── pixabay/                   # 337 images (84.2%)
│   ├── pexels/                    # 0 images
│   ├── unsplash/                  # 0 images
│   └── kaggle/                    # 0 images
│
├── api_downloads/                 # Categorized images
│   ├── back_view/                 # 187 images (target: 250)
│   ├── good_form/                 # Linked to form_quality_classifier
│   ├── amateur_poor_form/         # 9 images (target: 500)
│   └── metadata/                  # API metadata JSON files
│
├── form_quality_classifier/       # Form quality categories
│   ├── excellent_form/            # 300 images
│   ├── good_form/                 # 28 images (target: 228)
│   ├── needs_work/                # 15 images (target: 500)
│   └── poor_form/                 # 9 images (target: 500)
│
├── metadata/                      # Collection metadata
│   ├── pixabay_metadata.json      # Pixabay collection data
│   ├── pexels_metadata.json       # (will be created)
│   ├── unsplash_metadata.json     # (will be created)
│   └── kaggle_metadata.json       # (will be created)
│
├── logs/                          # Collection logs
│   ├── collection_*.log           # Real-time collection logs
│   └── collection_summary_*.json # Summary reports
│
└── scripts/                       # Collection scripts
    ├── collect_basketball_images.py
    ├── fast_parallel_collector.py
    ├── monitor_collection.py
    ├── categorize_images.py
    └── api_credentials.py
```

⚙️ Quality Filters

Image Resolution

- **Pixabay:** Minimum 1920x1080 (Full HD)
- **Pexels:** Minimum 1280x720 (HD)
- **Unsplash:** Minimum 1080p (one dimension)
- **Kaggle:** Minimum 640x480 (VGA)

Image Type

- **Format:** JPEG preferred (PNG acceptable)
- **Orientation:** Landscape preferred for most
- **Content:** Photos only (no illustrations/vectors)

Metadata Captured

- **Source:** API name, image ID

- **Query:** Search term used
- **Dimensions:** Width x height
- **Quality Indicators:** Likes, views, downloads
- **Attribution:** Photographer/user information
- **Timestamp:** Download datetime

🚀 Execution Timeline

Phase 1: Collection (In Progress)

- **Started:** December 13, 2025 16:30:10
- **Current Status:** Pixabay collection at 84.2%
- **Estimated Time:**
- Pixabay: ~10 minutes remaining
- Pexels: ~2-3 hours (due to 18s rate limit)
- Unsplash: ~8-10 hours (due to 72s rate limit)
- Kaggle: ~30-45 minutes
- **Total Estimated:** 10-14 hours for full collection

Phase 2: Categorization (Pending)

- **Trigger:** After collection completes or reaches 1,500+ images
- **Script:** `categorize_images.py`
- **Duration:** ~15-20 minutes
- **Output:** Images distributed to target categories

Phase 3: Verification (Pending)

- **Tasks:**
- Verify all targets met
- Remove duplicates
- Quality check sample images
- Generate final report
- **Duration:** ~10-15 minutes

Phase 4: RoboFlow Upload (Pending)

- **Prerequisites:**
- All categorization complete
- Quality verification passed
- Final counts match targets
- **Duration:** Depends on upload speed (~30-60 minutes for 1,500 images)

📈 Success Metrics

Collection Success

- ✅ **Quantity:** 1,500+ images collected

- ✅ **Quality:** High-resolution (1080p+) images
- ✅ **Diversity:** Multiple sources (4 APIs)
- ✅ **Metadata:** Complete tracking for all images

Categorization Success

- ✅ **Back View:** 200+ images (250 target)
- ✅ **Good Form:** 228+ images
- ✅ **Needs Work:** 500 images
- ✅ **Poor Form:** 500 images
- ✅ **Total:** 1,478 categorized images

Quality Indicators

- **Resolution:** 95%+ images at 1080p or higher
- **Relevance:** 90%+ images show basketball shooting
- **Variety:** Images from 4 different sources
- **Attribution:** 100% metadata captured

🔍 Monitoring Commands

Check Collection Progress

```
cd /home/ubuntu/basketball_app/training_data
python3 monitor_collection.py
```

Watch Progress Continuously

```
python3 monitor_collection.py --watch 30  # Updates every 30 seconds
```

View Collection Logs

```
tail -f logs/collection_*.log
```

Check Process Status

```
ps aux | grep collect_basketball_images.py
```

View Latest Metadata

```
cat metadata/pixabay_metadata.json | jq .
```

🎨 Image Attribution

Pixabay

- **License:** Content License (royalty-free)
- **Attribution:** Show “Images from Pixabay” when displaying
- **Commercial Use:** Allowed
- **Modifications:** Allowed

Pexels

- **License:** Pexels License (royalty-free)
- **Attribution:** Not required but appreciated
- **Commercial Use:** Allowed
- **Modifications:** Allowed

Unsplash

- **License:** Unsplash License (royalty-free)
- **Attribution:** Required (photographer name + Unsplash link)
- **Commercial Use:** Allowed
- **Modifications:** Allowed

Kaggle

- **License:** Varies by dataset (check individual dataset licenses)
- **Attribution:** As specified in dataset documentation
- **Commercial Use:** Check dataset terms
- **Modifications:** Check dataset terms

📝 Next Steps

Immediate (Automated)

1. ✅ Pixabay collection continues (84.2% complete)
2. ⏳ Pexels collection starts after Pixabay
3. ⏳ Unsplash collection starts after Pexels
4. ⏳ Kaggle collection starts after Unsplash

After Collection (Manual Trigger)

1. Run categorization script: `python3 categorize_images.py`
2. Verify target achievement
3. Review sample images for quality
4. Generate final report

Final Preparation

1. Remove duplicate images
2. Verify all metadata files
3. Prepare upload manifest for RoboFlow
4. Document collection summary

✅ Completion Checklist

Collection Phase

- [x] Extract Pixabay API key from screenshot
- [x] Set up API credentials for all 4 services
- [x] Create collection scripts with rate limiting
- [x] Create progress monitoring tools
- [🔄] Collect 400+ images from Pixabay (84.2%)
- [ ] Collect 400+ images from Pexels
- [ ] Collect 400+ images from Unsplash
- [ ] Collect 300+ images from Kaggle
- [ ] Total 1,500+ images collected

Categorization Phase

- [ ] Run automatic categorization
- [ ] Distribute uncategorized images
- [ ] Verify back view target (250 images)
- [ ] Verify good form target (228 images)
- [ ] Verify needs work target (500 images)
- [ ] Verify poor form target (500 images)

Quality Assurance

- [ ] Remove duplicate images
- [ ] Verify image quality (resolution, content)
- [ ] Validate metadata completeness
- [ ] Generate final collection report

RoboFlow Preparation

- [ ] Create upload manifest
- [ ] Prepare category labels
- [ ] Document dataset structure
- [ ] Ready for RoboFlow upload

Last Updated: December 13, 2025 16:33:38
Collection Status: 🔄 In Progress (22.5% complete)
Next Milestone: Complete Pixabay collection (~63 images remaining)